

## RESEARCH ARTICLE

# Exploring Predictive Variables Affecting the Sales of Companies Listed With Korean Stock Indices Through Machine Learning Analysis

GWANGSU LEE 

Korea SMEs and Startups Institute, Seoul 07074, South Korea

e-mail: 73gslee@gmail.com


**ABSTRACT** This study uses machine learning algorithms to explore predictor variables that determine whether the national statistical indices managed and announced by the Korean government influence the sales of companies listed on the Korea Composite Stock Price Index (KOSPI) and Korean Securities Dealers Automated Quotation (KOSDAQ). Further, it proposes a machine learning algorithm suitable for forecasting the sales of these companies. The sales of 1,470 companies listed on KOSPI and KOSDAQ with more than 20 years of history and 58 national statistical indices were analyzed. The predictor variables and performance were explored using the analysis data from 2000 to 2021 and the following machine learning algorithms: random forest, gradient boost, extreme gradient boosting, adaptive boosting, and categorical boosting. The analysis result confirmed that the national statistical indices contain different variables that affect the sales of listed companies by industry. The primary variable that showed the greatest influence in each industry was the industrial accident rate for manufacturing, finance and insurance, gold for construction, number of automobiles produced for wholesale and retail, and foreign exchange reserves for information and communication. The regression performance evaluation indicators—mean absolute error, mean squared error, and root mean squared error—were used to determine the optimal machine learning algorithm. The results showed that gradient boost achieved the best performance. Consequently, this study proposes using national statistical indices for companies to establish management strategies based on machine learning results.

**INDEX TERMS** Economic indices, machine learning, national statistical indices, predictor variables, sales.

## I. INTRODUCTION

Since the fourth industrial revolution was mentioned at the World Economic Forum in 2016, technological innovations have progressed in various fields through the fusion of artificial intelligence (AI), big-data analysis, and the Internet of Things with information and communication technology. Following this trend, the Korean government emphasizes data, networks, and AI. It has created an industrial ecosystem by introducing the Data Industry Act, amending three data laws, and AI ethics standards.

As a part of creating the industrial ecosystem, the Korean government has constructed and operated web-based

The associate editor coordinating the review of this manuscript and approving it for publication was Cheng Chin .

statistical information systems, the Statistics Korea e-Nara Index and Bank of Korea Economic Statistical System. These systems help people understand the social and economic situation at one place using 743 national statistical indices created based on indicators managed by 41 central administrative agencies and through private statistical data, such as price-, employment-, and production-related indices.

Price is the overall level that averages the prices of individual products traded in the market, considering their importance in economic life. The consumer price index is a representative index that refers to a price index that measures the average cost of living of urban households or changes in the purchasing power of currency by examining the price fluctuations of goods and services that consumers purchase in their daily lives [1], [2], [3].

Employment refers to a state in which one party provides labor to the other party, and the other party pays remuneration for it. The employment rate is a representative index that refers to the employed proportion of the population at a specific time among the working-age population (ages 15–64) [3]. Production refers to the process of converting raw materials into products or services through the input of people and equipment. The service index is a representative index that comprehensively represents the production activities of the entire service industry and individual industries. This index is determined by applying a weight representing the relative importance of each industry [3].

The Korean government has attempted to represent social and economic situations using national statistical indices. However, government press releases and news announced that the social and economic fields were directly or indirectly affected by COVID-19 after its outbreak in the second half of 2019. Consequently, problems with the national statistical indices provided through the Statistics Korea e-Nara Index and Bank of Korea Economic Statistical System have emerged.

Because the national statistical indicators representing social and economic conditions only provide statistical figures for indicators, only the changes in figures due to the influence of COVID-19 can be known via the indicators, but information on the related industries and companies affected by the national statistical indicators is not provided. This is the biggest problem in the fundamental purpose of using the national statistical index to reflect the social and economic situation. The contents of the national statistical indices provided by the Statistics Korea e-Nara Index and Bank of Korea Economic Statistical System are as follows: First, national statistical indices only provide monthly, quarterly, and yearly statistics; therefore, we can only know the changes from the previous month, quarter, and year. Second, a comparative analysis cannot be performed on the same statistical cycle for every national statistical index because not all national statistical indices provide monthly, quarterly, and annual statistical values. Third, most national statistical indices do not provide related indices except for the representative national statistical indices. Hence, when the statistical values of national statistical indices increase or decrease, the related indices will be affected and cannot be determined.

As such, national statistical indicators are at the level of providing only one-dimensional information about indicators. Now it is time to improve national statistical indicators. Rather than simply presenting national statistical indicators, information on the impact of national statistical indicators on corporate sales according to corporate characteristics should be presented. However, because research on this is insufficient, knowing what kind of relationship exists between national statistical indicators and corporate characteristics is not possible; therefore, there will be limitations in presenting them.

National statistical indicators are believed to affect different types of industries. However, the research related to this is insufficient, or no specific research data exist yet.

A previous study analyzed the relationship between national statistical indicators and one company; however, to date, no study has targeted multiple companies or classified companies by industry. Research methodology also uses machine learning algorithms to suggest optimal algorithms suitable for data characteristics. Because machine learning algorithms can have different predictive performance due to differences in analysis data, dependent variable type, and number of predictors, the research subject is not a company or research results targeting one company cannot be generalized. Thus, I hypothesize that the lack of related studies examining the relationship between national statistical indicators and companies is also affecting [12].

The Korea Composite Stock Price Index (KOSPI) and Korean Securities Dealers Automated Quotation (KOSDAQ) are representative stock indices that represent the economic situation of Korea and best reflect political, social, and economic factors and corporate management performance. KOSPI represents the flow of all stocks (excluding KOSDAQ) listed on the Korea Stock Exchange. Companies with an equity capital of more than 30 billion won are listed as part of corporate size requirements. KOSDAQ is a stock market operated by the KOSDAQ committee for small- and medium-sized enterprises and venture companies with a function similar to that of the US NASDAQ. Among the company size requirements, companies with an equity capital of 1.5 billion won or more or market capitalization of 9 billion won or more are listed. As such, companies registered with the KOSPI and KOSDAQ markets reflect the Korean economic situation as a stock index.

This study selected KOSPI- and KOSDAQ-listed companies as companies that are affected by social and economic conditions to understand whether national statistical indicators affect companies.

In addition, KOSPI- and KOSDAQ-listed companies were classified according to the type of business, which is a characteristic of the company, to determine whether the index affecting the company's sales varies with the type of business, in which the national statistical index is a characteristic of the company. The research methodology also used machine learning algorithms to analyze national statistical indicators and real data produced by companies without using traditional statistical techniques such as surveys and focus group interviews (FGIs). The performance of each machine learning algorithm was evaluated and a machine learning algorithm suitable for predicting sales of KOSPI- and KOSDAQ-listed companies was presented. In addition, by studying companies that have not yet been specifically studied, national statistical indicators suitable for corporate characteristics were presented to enhance corporate management activities.

This study contributes to the field in three aspects because of deriving the factors that affect the sales of KOSPI- and

KOSDAQ-listed companies from the national statistical indicators of the Korean government.

First, the subject of research using macroeconomic indicators was expanded to industries that are characteristic of companies. Most of the related studies analyzed the correlation between macroeconomic indicators, and studies analyzing the correlation between macroeconomic indicators and one company are insufficient or have not been studied in detail yet. Therefore, there were limitations in presenting the relationship between macroeconomic indicators and companies. To improve this, the research target was expanded to the industry, which is the characteristic of a company, and the cornerstone was laid for a study that analyzes the relationship between macroeconomic indicators and companies.

Second, the national statistical indicators were suggested to have different indicators affecting the sales of listed companies depending on the type of business, which is the characteristic of the company. Related studies have a limitation in presenting the macroeconomic indicators affecting each industry, because no data exist specifically on the relationship between macroeconomic indicators and industry, which is a characteristic of companies. To improve this, for the first time, an empirical analysis was conducted using machine learning algorithms for national statistical indicators and business characteristics, and standards for which national statistical indicators should be used as basic data for companies to establish management strategies were established.

Third, the optimal machine learning algorithm was suggested to vary depending on the sales of listed companies and sales of listed companies by industry, which are data characteristics. Related studies have suggested optimal machine learning algorithms suitable for analysis data, but there is a limit to suggesting that the optimal machine learning algorithm varies with the same analysis data characteristics when analyzed comprehensively. To improve this, the optimal machine learning algorithm was found to vary because of analyzing the sales of listed companies and sales of listed companies by industry according to the characteristics of the analysis data. These results found that differences in analysis data, dependent variable types, and number of predictors change not only the predictability of machine learning but also the machine learning algorithm.

The remainder of this paper is organized as follows: Section II discusses the relevant literature. Section III introduces the variable explanation and analysis used in the study, and Section IV presents the machine learning analysis and comparison results. Section V concludes the study and presents directions for future research.

## II. RELATED WORK

Studies on macroeconomic indicators of social and economic conditions have been continuously conducted. Among them, studies investigated the effects of macroeconomic indicators on the stock market, and studies using machine and deep learning algorithms to predict economic trends and stock markets using economic and financial variables

are in progress [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15].

According to Sova and Lukianenko [4], stock indices in developed countries are affected by interest rates and monetary policies. However, stock indices in developing countries are affected from a long-term rather than short-term perspective [4]. Fernandez and Li [5] published a report, which suggested that consumer activity and monetary policy affect the Philippine stock market; it is affected by all macroeconomic variables except interest rates.

Guven et al. [6] suggested that the number of daily deaths and daily increase in patients infected with COVID-19 had a negative effect on the stock index and that the response policy of the government had a positive impact on the stock index.

Loang and Ahmad [7] suggested that before the COVID-19 pandemic, company information (company size, return on assets (ROA), return on equity (ROE), earnings per share (EPS)) and macroeconomic variables (export rate, import rate, real GDP, nominal GDP, FDI, IPI, and unemployment rate) affected stock indices in the US and China. However, after the COVID-19 pandemic, company information had no impact, and only macroeconomic variables affected stock indices.

Liu et al. [8] examined the relationships between the consumer and producer price indices with the future prices of the China CSI 300 stock index. They suggested that the consumer and producer price indices affected the futures prices of the CSI 300 stock index. Table 1 summarizes the studies in which macroeconomic indicators affect the stock index.

I reviewed previous studies that applied optimal machine and deep learning algorithms to predict the future using macroeconomic indices. Singh [9] analyzed data spanning 25 years from April 22, 1996, to April 16, 2021, using adaptive boosting ((AdaBoost), k-nearest neighbors, linear regression, artificial neural network, random forest, stochastic gradient descent, support vector machine, and decision tree algorithms to derive optimal machine and deep learning algorithms to predict the Nifty 50 index of the Indian market. Consequently, he suggested stochastic gradient descent as the optimal algorithm. Bhardwaj et al. [10] used an artificial deep neural network, random forest, gradient boost, ridge regression, and k-nearest neighbors algorithms to predict per capita GDP using 262 growth indicators, development, health, energy, and finance in 33 OECD countries. After analyzing data spanning 22 years from 1996 to 2017, they determined the artificial deep neural network as the optimal algorithm for predicting per capita GDP.

Chatterjee et al. [11] analyzed stock price data from January 2004 to December 2019 using autoregressive integrated moving average (ARIMA), random forest, multivariate adaptive regression splines (MARS), recurrent neural network, and long short-term memory algorithms to determine the optimal algorithm for predicting the future prices of three major stocks on the National Stock Exchange of India. Consequently, they recommended the MARS algorithm as the optimal algorithm for predicting future prices. Lee [12]

**TABLE 1. Studies on the influence of macroeconomic indicators on the stock index.**

Researcher	Research content	Research variables
Sova & Lukianenko [4]	Analyze, using the vector autoregressive model, whether monetary policy affects the stock market in developed and developing countries	Interest rate and monetary
Fernandez & Li [5]	Analyze whether macroeconomic variables of business activity, consumer activity, and monetary policy affect the Philippine stock market	Interest rate, monetary, remittances, consumer spending, and industrial production index
Guyen <i>et al.</i> [6]	Analyze whether the daily death toll, daily patient growth rate, and government response policies affected the stock market due to the impact of COVID-19	The daily death toll, daily patient growth rate, and government response policy
Loang & Ahmad [7]	Analyze whether company information and macroeconomic variables before and after COVID-19 affect the stock markets of the United States and China	Firm size, return on asset (ROA), return of equity (ROE), earning per share (EPS), export rate, import rate, real GDP, nominal GDP, FDI, IPI, and unemployment rate
Liu <i>et al.</i> [8]	Analyze whether the consumer price index and producer price index affect the futures price of China's CSI 300 stock index	Consumer price index, producer price index

examined the variable that most significantly affects the sales of printing-related small and medium enterprises (SMEs) among 22 economic statistical indices related to prices, growth, employment, and interest rates. To determine the optimal algorithm, they analyzed the sales of printing-related SMEs from August 2013 to November 2021 using the random forest, extreme gradient boosting (XGBoost), and light gradient boosting machine algorithms (LightGBM). They presented the consumer price index and the cost-of-living index for living necessities, which are price indices, as essential variables that affect the sales of printing-related SMEs. They also found random forest (RF) as the optimal algorithm. Moreover, Lee [13] explored the index that most influenced drugstore sales among the 28 government statistical indices related to price, economy, employment, and interest rate and analyzed the sales of drugstores from January 2016 to December 2021 using RF, extreme gradient boosting, light

gradient boosting machine, and categorical boosting (CatBoost) algorithms to determine the optimal algorithm. They found the economic sentiment index, cyclical component of the coincidence index, and consumer sentiment index as the variables influencing drugstore sales. They found RF as the optimal algorithm.

Gaspareniene *et al.* [14] used the decision tree, RF, linear regression, extreme gradient boosting, feedforward neural network, recurrent neural network, and long short-term memory algorithms to predict the S&P 500 index using 27 indicators related to macroeconomic, labor market, real estate market, credit market, and money supply. Results showed that the treasury bill, crude oil price, and personal savings were the important variables affecting the S&P 500 index, and RF was presented as the optimal algorithm.

Bharat *et al.* [15] used the linear regression, classification and regression trees (CART), generalized linear model (GLM), ARIMA, autoregressive moving average with exogenous (ARMAX), and vector autoregressive (VAR) algorithms to predict crude oil prices using macroeconomic indicators such as GDP, interest rates, investment, unemployment rate, US dollar exchange rate, and consumer preference index. The results of the analysis showed that the US dollar exchange rate and consumer preference index were the important variables affecting the price of crude oil, and ARMAX was the optimal algorithm. Table 2 summarizes the studies that used machine learning algorithms and macroeconomic indicators.

Studies in which macroeconomic indicators affect the stock index commonly present economic indicators that affect the stock index the most among macroeconomic indicators. Studies predicting economic indicators through machine learning analysis commonly present economic indicators that affect dependent variables among macroeconomic indicators, and studies comparing and analyzing multi-machine learning algorithms present optimal machine learning algorithms suitable for analysis data characteristics.

As such, related studies use macroeconomic indicators to present those that most affect stock index, GDP per capita, sales of printing companies, pharmacy sales, and crude oil prices and explain a significant relationship with dependent variables. However, after analyzing related studies, limitations of the study were found. Macroeconomic variables that affect dependent variables are all different. This may be a natural result because the independent and dependent variables used in the study are different. However, according to the studies of Lee [12] and Lee [13], independent variables commonly used national statistical indicators; additionally, only dependent variables were analyzed by dividing them into printing companies and pharmacies among small businesses, and macroeconomic variables that affected them were different. Printing companies and pharmacies were classified into manufacturing, wholesale, and retail industries, and the research results were presented differently depending on the industry.

In other words, related studies only explain the relationship between macroeconomic indicators and dependent variables,



**TABLE 2. Economic indicator prediction research through machine learning analysis.**

Researcher	Research content	Analysis method
Singh [9]	Predicts India's Nifty 50 index with machine learning and presents an optimal machine learning algorithm	Adaptive boosting, k-nearest neighbors, linear regression, artificial neural network, RF, stochastic gradient descent, support vector machine, and decision trees
Bhardwaj <i>et al.</i> [10]	Through machine learning analysis, predict GDP per capita in 33 OECD countries using 262 indicators related to growth, development, health, energy, and finance.	Artificial deep neural network, RF, gradient boost, ridge regression, and k-nearest neighbors
Chatterjee <i>et al.</i> [11]	Predict the future price of stocks on the National Stock Exchange of India using machine learning and determine an optimal machine learning algorithm.	ARIMA, RF, MARS, RNN, and LSTM
Lee [12]	Through machine learning analysis, among 22 economic indicators related to prices, growth, employment, and interest rates, the variables that most affect the sales of printing-related SMEs and the optimal machine learning algorithm are determined.	RF, extreme gradient boosting, and light gradient boosting machine
Lee [13]	Through machine learning analysis, among 28 government statistical indicators related to prices, economy, employment, and interest rates, variables that most affect pharmacy sales and optimal machine learning algorithms are determined	RF, extreme gradient boosting, light gradient boosting machine, and categorical boosting
Gaspareniene <i>et al.</i> [14]	Through machine learning analysis, 27 indicators related to US economic indicators, such as general macroeconomic indicators, labor market indicators, real estate market indicators, credit market indicators, and monetary supply indicators, are used to predict the S&P 500 index. The optimal machine learning algorithm is determined.	Decision tree, RF, linear regression, extreme gradient boosting, feedforward neural network, recurrent neural network, and long short-term memory
Bharat <i>et al.</i> [15]	Through machine learning analysis, macroeconomic indicators such as GDP, interest rate, investment, unemployment rate, dollar exchange rate, and consumer preference index are used to determine the variables that most affect crude oil prices and the optimal machine learning algorithm.	Linear regression, CART, GLM, ARIMA, ARMAX, and VAR

but there is a limit to the detailed analysis on how macroeconomic variables affect dependent variables according to related industries or business characteristics. Owing to these limitations, it is not known whether macroeconomic indicators that represent social and economic conditions affect companies and related industries, whether they affect companies in common depending on corporate characteristics, or whether macroeconomic indicators that affect them vary depending on corporate characteristics.

Accordingly, this study analyzes whether macroeconomic indicators affect corporate characteristics, identifies which indicators affect corporate sales the most according to corporate characteristics, and proposes an algorithm suitable for predicting corporate sales.

### III. RESEARCH METHOD

#### A. DATASET

To explore whether national statistical indices representing social and economic conditions affect the sales of KOSPI- and KOSDAQ-listed companies, 58 statistical indices that were compiled annually among 743 national statistical indices were selected as features. The 58 statistical indicators consisted of 6 traffic infrastructure-related, 14 growth-related, 4 leisure-related, 5 price-related, 4 customs-related, 6 production-related, 4 employment-related, 3 interest-rate-related, 3 population-related, and 6 other indicators. Table 3

shows the variable and basic statistics for the feature dataset by category for the 58 statistical indices.

As the dependent variable, sales of KOSPI- and KOSDAQ-listed companies were selected. For companies listed on the KOSPI and KOSDAQ, the annual sales of 1,470 KOSPI- and KOSDAQ-listed companies with more than 20 years of business among those registered with the Korea Listed Companies Association were selected. The 1,470 KOSPI- and KOSDAQ-listed companies consisted of 998 manufacturing, 59 construction, 125 wholesale and retail, 159 information and communication, and 129 finance and insurance companies. Table 4 shows the number and basic statistics of the 1,470 companies by industry.

#### B. MACHINE LEARNING ANALYSIS

Python (version 3.9.7) was used for the machine learning analysis. In addition, machine learning algorithms such as RF, gradient boost, XGBoost, AdaBoost, and CatBoost were used for regression analysis.

Ensemble learning refers to a technique that produces multiple classifiers. It combines their predictions to derive accurate final predictions, and the types of ensemble learning are divided into voting, bagging, and boosting. In voting and bagging, multiple classifiers determine the final prediction result through voting. Voting refers to combining classifiers

**TABLE 3. Descriptions of variables and basic statistics for the feature dataset.**

Category	Feature	Description	Min	Max	Mean	Std. Dev.
Transportation infrastructure	inpa1	Number of international passengers	3235646	90900322	41953137.22	23769652.99
	dopa2	Number of domestic passengers	16847870	33386561	23495917.09	5661833.93
	incv3	International cargo volume	1863832	4168808	3072002.86	658567.93
	docv4	Domestic cargo volume	181785	434228	310437.22	75679.06
	inro42	Number of international routes	204	378	310.45	53.13
	doro43	Number of domestic routes	19	37	23.77	4.98
Growth	exch5	Won-dollar exchange rate	929.80	1313.50	1127.42	98.93
	defa6	Number of bill defaulters	183	6693	2184.77	1920.88
	kosp7	KOSPI Index	504.62	2977.65	1699.85	692.85
	kosd8	KOSDAQ Index	332.05	1033.98	608.58	174.39
	cons13	Construction investment amount	178.6	282.9	230.95	26.67
	foex34	Foreign exchange reserves	96198117	463118362	288167747.45	110957256.09
	lapr48	Labor productivity	52.5	120.5	91.85	21.20
	pofa49	Policy funding amount	17737	62900	37798.59	12741.57
	pofc50	Number of policy-funded companies	15197	59968	24577.05	12072.41
	aupr51	Number of automobiles produced	2946	4657	3872.91	529.80
	audd52	Number of domestic automobiles sold	1094	1885	1523.32	257.73
	auca53	Automobile export amount	153.2	497	381.30	109.81
	auia54	Automobile import amount	18.4	142.5	71.62	37.98
	faii63	Estimated index of equipment investment	57.2	122.7	86.4	20.17
Leisure	toin9	Tourism revenue	53.4	207.5	108.84	47.04
	toex10	Tourist expenditure	61.7	315.3	160.5	70.94
	foto11	Number of foreign tourists visiting Korea	967	17503	8908.14	4706.47
	trav12	Number of overseas travelers	122.3	2871.4	1327.63	769.97
Price	prod25	Producers price index	77.76	109.6	94.97	10.59
	livp26	Cost-of-living index	63.15	102.5	85.34	12.58
	comp27	Consumer price index	63.15	102.5	85.34	12.58
	impp29	Import price index	74.39	138.88	103.36	20.37
	expp32	Export price index	94.74	129.71	111.27	10.20
Custom import and export	impa28	Import amount index	32.62	142.63	89.12	33.60
	impv30	Import volume index	47.19	124.32	84.58	22.36
	expa31	Export amount index	28.91	128.07	81.21	31.17
	expv33	Export volume index	26.14	120.48	74.63	31.63
Production	mapr35	Manufacturing production index	49.39	114.3	85.34	21.25
	main36	Manufacturing inventory index	47.42	122.4	80.83	24.38
	mafo37	Manufacturing shipment index	52.51	106	85.46	18.53

**TABLE 3. (Continued.) Descriptions of variables and basic statistics for the feature dataset.**

	maui68	Manufacturing utilization rate index	953	10848	103.1	3.91
	whre61	Wholesale and retail index	44	1224	83.66	23.95
	sipi62	Index of service	44	1224	83.66	23.95
Employment	empr55	Employment rate	58.5	60.9	59.9	0.69
	uner56	Unemployment rate	3.1	4.4	3.61	0.31
	empp57	Number of employed	21173	27273	24505.05	1949.07
	ecap58	Economically active population	22151	28310	25420.41	2021.25
Working conditions	accr44	Industrial accident rate	0.48	0.9	0.66	0.12
	deto45	Occupational disaster death toll	1777	2748	2145.05	305.11
	dico46	Number of people with occupational disease	4051	20435	9408.86	3732.48
Interest rate	lera65	Lending interest rate	2.8	8.55	5.22	1.62
	rera66	Receiving rate	1.05	7.01	3.32	1.63
	bara69	Base rate	0.5	5.25	2.75	1.36
Population	estp38	Population estimate	47008111	51836239	49684486.59	1597310.61
	birt39	Combined fertility rate	0.81	1.48	1.15	0.16
	move22	Number of people moving in the population	7104	9584	8227.64	847.58
Consumption	ausi67	Automobile sales index	354	135.8	72.55	32.99
	resi64	Retail sales index	453	127	84.19	24.72
Raw materials	oilp23	International oil price Dubai	18.28	107.89	58.91	28.10
	gold24	Gold	272.25	1898.36	1036.75	511.90
Stability	exbo59	External bonds	156672.5	1080325.1	546084.06	294512.29
	exde60	External debts	114666.1	632393.6	333227.5	146534.68

**TABLE 4. Number of listed companies by industry and basic statistics.**

Industry	Number of listed companies	Min	Max	Mean	Std. Dev.
Manufacturing	998	278578	819145	520350.55	174194.53
Construction	59	462145	1529370	1018106.09	387503.82
Wholesale and retail	125	481677	1052589	812377.95	152828.09
Information and communication	159	240028	494054	355375.5	89491.04
Finance and insurance	129	612054	2342780	1480685.64	599796.76
Total	1470	376188	888490	657708.27	201740.76

with different algorithms; with bagging, all classifiers are based on the same type of algorithm, but learning is performed with different data sampling. Boosting is a learning method that reduces errors by weighting incorrectly predicted

data while sequentially learning and predicting weak learners.

A representative bagging method, RF, achieves high accuracy by minimizing prediction errors and overfitting by

**TABLE 5. Packages by the machine learning algorithm.**

Algorithms	Packages
Random forest	RandomForestRegressor
Gradient boost	GradientBoostingRegressor
XGBoost	XGBRegressor
AdaBoost	AdaBoostRegressor
CatBoost	CatBoostRegressor

maximizing randomness in sample and variable selection to improve predictive power and reduce overfitting, which are the problems with decision trees.

Boosting methods include AdaBoost, gradient boost, XGBoost, and CatBoost. AdaBoost is a method of boosting while weighting error data and is similar to AdaBoost but uses gradient descent to update weights.

XGBoost alleviates the problems of gradient boost such as slow execution time and lack of regularization. It reduces the execution time with parallel processing support, has a strong durability against overfitting owing to the addition of the regularization function, and has an optimized number of repetitions with its built-in cross-validation function, suggesting excellent prediction performance [16].

CatBoost is based on gradient boost. However, gradient-boost-based algorithms have the problem of one-hot encoding, which rapidly increases the number of variables and slows the learning speed of the model when using categorical variables. CatBoost was created to address this issue, exhibiting excellent performance when analyzing datasets composed of categorical variables. Table 5 lists the packages according to the machine learning algorithms.

The analysis procedure involved data preprocessing, model learning, and model evaluation, as shown in Fig. 1. A detailed examination of each step is as follows.

Data preprocessing was performed before model learning in the sequence of data matching, cleaning, and transformation, as follows: First, 58 statistical indicators used as features in the data matching step and 1,470 listed companies' sales data layouts used as targets were analyzed. Looking at Fig. 2, features are composed of years and indicator values, and targets are composed of years and sales. Because there is no unique identification information in features and targets but only the year, which is a common variable that exists, data were combined based on the year using nonparametric matching, a statistical matching method among the data matching methods [17], [18], [19].

Data matching is a technique for combining different datasets. It is divided into exact and statistical matching depending on the existence or absence of data combinations using unique identifiers such as resident registration, business registration, passport, and license numbers. Exact matching combines data having the same identifier value for datasets having unique identifiers. Statistical matching combines data

when a dataset does not have unique identifiers [13], [17], [19], [20].

Statistical matching is divided into non-parametric and parametric estimation matching, depending on the existence of data combinations using common and unique variables in datasets. Non-parametric estimation matching combines data using only common variables in each dataset. In contrast, parametric estimation matching combines data using only common and unique variables in each dataset [13], [18], [19].

Second, in the data cleaning step, features had different management and presentation timings according to indicators, and indicators that had been manually managed before the establishment of the information system did not have yearly indicators. The target also had different start-up dates, and different sales reporting years existed depending on the company's business history based on the start-up date. For accurate data analysis, data that did not have yearly indicators among features were deleted, and the target refined the data to the common reporting year among the different reporting years according to the company's performance. Based on the refined features and target data, the data analysis period was set from 2000 to 2021, the period in which the years existed in common.

Third, in the data transformation step, features and targets showed different units and ranges of indicator values according to indicators. In particular, in the case of features, the impact of the target was identified by the unit and range of the indicator value on the same basis to determine the importance and performance evaluation of variables between features. To solve this problem, all variables were converted into log values to make them into a normal distribution for feature scaling, which is an operation for adjusting the index values and ranges of features and targets to a certain level.

Model learning is the process that determines the best performance evaluation index based on the analysis data, and it is performed in the following sequence:

First, the analysis data are divided into training and testing data. The reason is that testing data are needed to evaluate how well the model has learned from the training data. In addition, 80% is set as training data and 20% as testing data through random functions to prevent sampling bias of training and testing data. Generally, the performance for time-series data is evaluated by using past data as training data and the latest data as testing data. However, there is a problem of overfitting the testing data if verifying and correcting the model performance is repeated using fixed testing data. Cross-validation is performed to prevent this problem, which consists of training and evaluation after composing training and validation datasets of several different sets to remove data bias [12], [13].

In this study, 5 folds were used for cross-validation, the hyperparameter setting range for each machine learning algorithm was set as shown in Table 6, and the optimal hyperparameters were designated through GridSearchCV provided by Scikit-Learn within the specified range.



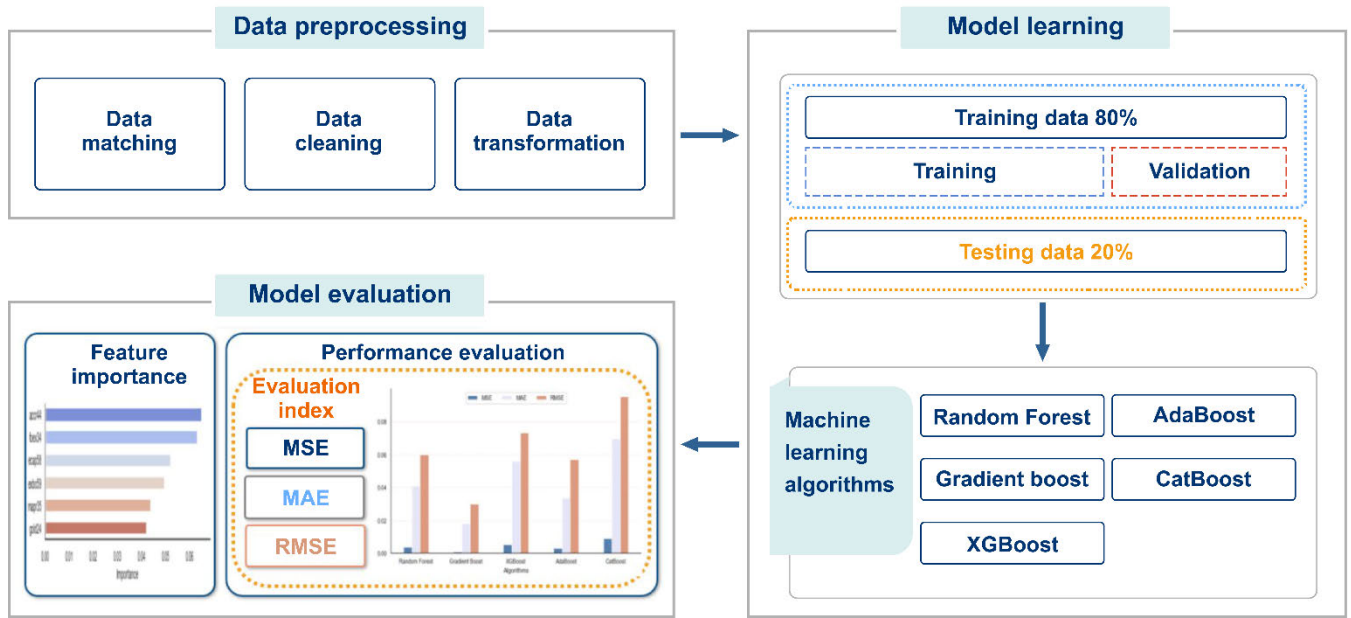


FIGURE 1. Machine learning analysis process.

Feature			Target		
Indices	Yyyy	Indices value	Industry	Yyyy	Sales
Import price index	2000	76.61	Manufacturing	2000	520,350
Export price index	2001	124.77	Construction	2001	1,018,106
Index of service	2002	53.3	Wholesale and retail	2002	812,377
⋮	⋮	⋮	⋮	⋮	⋮
Employment rate	2021	60.5	Finance and insurance	2021	1,480,685

FIGURE 2. Analysis data layout.

Second, regression analysis was performed to calculate the optimal performance evaluation indicators—mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE)—by applying machine learning algorithms such as RF, XGBoost, AdaBoost, AdaBoost, and CatBoost.

MAE was obtained by averaging the difference between the target value  $y_i$  and predicted value  $\bar{y}_i$  converted into absolute values as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}_i|. \quad (1)$$

The MSE was obtained by averaging the square of the differences between the target value  $y_i$  and predicted value  $\bar{y}_i$  as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2. \quad (2)$$

TABLE 6. Hyperparameter setting range.

Algorithms	Hyperparameter Range
Random forest	n_estimators: [100, 200, 300, 400, 500, 1000], max_depth: [6, 8, 10, 12], min_samples_leaf: [1, 2, 4, 8, 12, 18], min_samples_split: [2, 4, 8, 16, 20]
Gradient boost	n_estimators: [100, 200, 300, 400, 500, 1000], learning_rate: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5], max_depth: [2, 3, 4, 5, 6, 7, 8], min_samples_leaf: [1, 2, 4, 8, 12, 18], min_samples_split: [2, 4, 8, 16, 20], subsample: [0.5, 0.7, 0.8, 0.9, 1]
XGBoost	n_estimators: [100, 200, 300, 400, 500, 1000], learning_rate: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5], max_depth: [2, 3, 4, 5, 6, 7, 8], gamma: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 1], min_child_weight: [1, 2, 3, 4, 5], subsample: [0.5, 0.7, 0.8, 0.9, 1], colsample_bytree: [0.5, 0.6, 0.7, 0.8, 0.9, 1]
AdaBoost	n_estimators: [100, 200, 300, 400, 500, 1000], learning_rate: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5]
CatBoost	n_estimators: [100, 200, 300, 400, 500, 1000], learning_rate: [0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1], depth: [2, 3, 4, 5, 6, 7, 8], min_child_samples: [1, 2, 3, 4, 5], subsample: [0.5, 0.7, 0.8, 0.9, 1]

The RMSE is the square root of MSE as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2}. \quad (3)$$

The closer MAE, MSE, and RMSE are to zero, the more accurate the model and the better is the applied machine learning algorithm [9], [12], [13], [16].

The model evaluation stage involves comparatively analyzing the performance evaluation indices according to machine learning algorithms and deriving the importance of features that influence the target. This was performed sequentially as follows:

First, the optimal machine learning algorithm was determined by analyzing MAE, MSE, and RMSE, which are the metrics for evaluating the regression performance of the machine learning algorithm.

Second, the importance of the features influencing the target variable was derived using the optimal machine learning algorithm. The machine learning algorithm used in this study was based on a decision tree. Decision trees are machine learning models that predict target variables while continuously branching data into specific features, such as trees. When dividing nodes during training, the concept of Gini impurity is used for classification. The MSE is used for regression for branching; therefore, the information gain is the highest. The information gain is a general term for the performance gain obtained by branching a decision tree to a specific feature, and the feature importance is calculated based on this.

The feature importance is a number indicating how much the variable affects the target for each machine learning algorithm, and the greater the number, the higher the importance [12], [13], [21]. Specifically, through feature importance, the statistical index with the highest importance between the national statistical indices has the greatest predictive power for the sales of KOSPI- and KOSDAQ-listed companies.

Third, the performance evaluation index of machine learning algorithms was compared and analyzed by industry. The difference was obtained by deriving feature importance to understand whether the optimal machine learning algorithm and variables affecting sales by industry vary depending on the industry of the listed companies.

#### IV. RESULTS

##### A. RESULTS OF MACHINE LEARNING ANALYSIS OF THE SALES OF LISTED COMPANIES

MAE, MSE, and RMSE of the regression performance evaluation indices were analyzed for each machine learning algorithm using the data of 1,470 companies to determine the national statistical indices with the largest influence on the sales of KOSPI- and KOSDAQ-listed companies. The results are summarized in Table 7.

From the regression analysis results in Table 7, gradient boost showed a better regression performance than that of RF, XGBoost, AdaBoost, and CatBoost according to the evaluation criteria of the regression performance evaluation indicators, MAE, MSE, and RMSE. In other words, in predicting whether national statistical indicators affect the sales

TABLE 7. Regression-analysis results for the sales of listed companies by the machine learning algorithm.

Algorithms	MAE	MSE	RMSE	Time (s)
RF	0.0408	0.0036	0.0599	0.7656
Gradient boost	0.0181	0.0009	0.0299	0.1554
XGBoost	0.0561	0.0053	0.0731	0.0349
AdaBoost	0.0337	0.0032	0.0570	0.2787
CatBoost	0.0695	0.0090	0.0951	0.1437

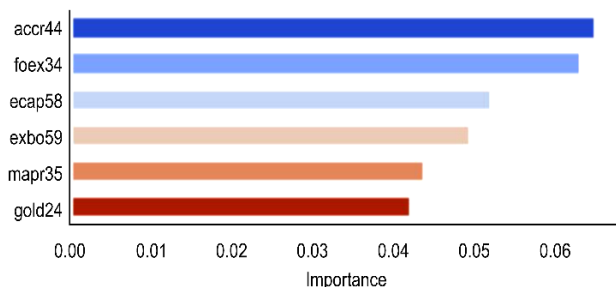


FIGURE 3. Feature importance Top 6.

of KOSPI- and KOSDAQ-listed companies, gradient boost can be considered the optimal machine learning algorithm.

The results of extracting the top six important features, which are the top 10% of the 58 independent variables, using gradient boost determined as the optimal machine learning algorithm, are shown in Fig. 3 and Table 8.

As for the feature importance presented in Fig. 3, Industrial accident rate, Foreign exchange reserves, Economically active population, External bonds, Manufacturing production index, and Gold are shown in order. Among them, the numerical value of Industrial accident rate was the largest, indicating that the predictive power of the sales of listed companies was the highest.

Furthermore, the features by category in Table 3 were examined concerning the feature importance results in Table 8. The sales of KOSPI- and KOSDAQ-listed companies were affected more by working conditions, growth, employment, stability, production, and raw materials than by prices, leisure, population, and transportation infrastructure.

##### B. MACHINE-LEARNING-ANALYSIS RESULTS FOR THE SALES OF LISTED COMPANIES BY INDUSTRY

Listed companies were classified into manufacturing, construction, wholesale and retail, information and communication, finance, and insurance according to industry to determine whether the optimal machine learning algorithm and variables affecting sales by industry varied with industry among the listed companies. Subsequently, the regression performance evaluation indices MAE, MSE, and RMSE were examined for each machine learning algorithm. The results are summarized in Table 9.

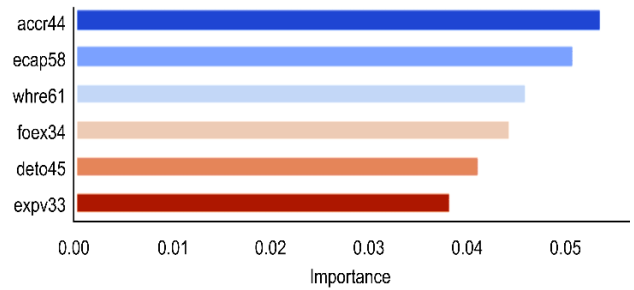
**TABLE 8. Top six feature importance of listed companies.**

Ranking	Category	Feature	Description
1	Working conditions	accr44	Industrial accident rate
2	Growth	foex34	Foreign exchange reserves
3	Employment	ecap58	Economically active population
4	Stability	exbo59	External bonds
5	Production	mapr35	Manufacturing production index
6	Raw materials	gold24	Gold

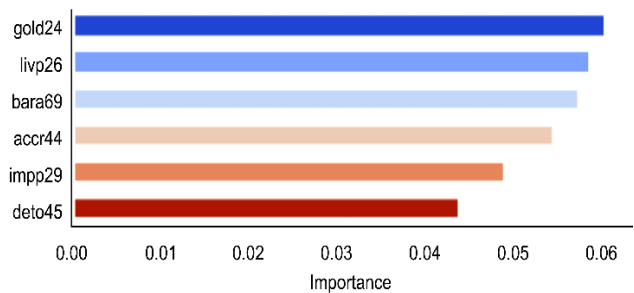
**TABLE 9. Regression analysis results for each machine learning algorithm by industry.**

Industry	Algorithms	MAE	MSE	RMSE	Time (s)
Manufacturing	RF	0.0363	0.0019	0.0434	0.7727
	Gradient boost	0.0160	0.0003	0.0180	0.1493
	XGBoost	0.0253	0.0010	0.0317	0.0352
	AdaBoost	0.0208	0.0013	0.0367	0.2669
	CatBoost	0.0560	0.0051	0.0717	0.1455
Construction	RF	0.0587	0.0038	0.0620	0.7779
	Gradient boost	0.0576	0.0036	0.0602	0.1549
	XGBoost	0.0653	0.0051	0.0715	0.0333
	AdaBoost	0.0707	0.0064	0.0802	0.2679
	CatBoost	0.1027	0.0156	0.1250	0.1422
Wholesale and retail	RF	0.0718	0.0084	0.0914	0.7770
	Gradient boost	0.0457	0.0030	0.0544	0.1555
	XGBoost	0.0474	0.0033	0.0576	0.0355
	AdaBoost	0.0560	0.0041	0.0638	0.3376
	CatBoost	0.0357	0.0020	0.0444	0.1450
Information and communication	RF	0.0413	0.0020	0.0449	0.7763
	Gradient boost	0.0254	0.0009	0.0308	0.1525
	XGBoost	0.0425	0.0020	0.0445	0.0343
	AdaBoost	0.0296	0.0011	0.0330	0.2675
	CatBoost	0.0636	0.0072	0.0850	0.1464
Finance and insurance	RF	0.0816	0.0112	0.1058	0.1019
	Gradient boost	0.1272	0.0323	0.1796	0.1553
	XGBoost	0.1645	0.0636	0.2523	0.0347
	AdaBoost	0.1082	0.0183	0.1354	0.0717
	CatBoost	0.1467	0.0299	0.1728	0.1896

Looking at the results of the regression analysis in Table 9, Gradient Boost showed the best regression performance in the manufacturing, construction, and information and communication industries according to the evaluation criteria of



**FIGURE 4. Feature importance Top 6 of the manufacturing industry.**



**FIGURE 5. Feature importance Top 6 of the construction industry.**

the regression performance evaluation indicators MAE, MSE, and RMSE. CatBoost showed the best regression performance in the wholesale and retail industries, whereas RF showed the best regression performance in the financial and insurance industries. In other words, it was confirmed that the optimal machine learning algorithm varied with the industry in predicting whether national statistical indicators affected sales by industry among KOSPI- and KOSDAQ-listed companies. Using the optimal machine learning algorithm by industry, the top 6 feature importance by industry, which is the top 10% of the 58 independent variables, was extracted, as shown in Figs. 4–8.

Fig. 4 shows the importance of features in the manufacturing industry. The descending order of feature importance is as follows: industrial accident rate, economically active population, wholesale and retail index, foreign exchange reserves, occupational disaster death toll, and export volume index. Among these, the industrial accident rate shows the highest predictive power for manufacturing company sales.

Fig. 5 shows the importance of features in the construction industry. The descending order of feature importance is as follows: gold, cost-of-living index, base rate, industrial accident rate, import price index, and occupational disaster death toll. Among these, gold shows the highest predictive power for construction company sales.

Fig. 6 shows the feature importance of the wholesale and retail industries. The descending order of feature importance is as follows: number of automobiles produced, cost-of-living index, export amount index, domestic cargo volume, export price index, and international cargo volume. Among these,

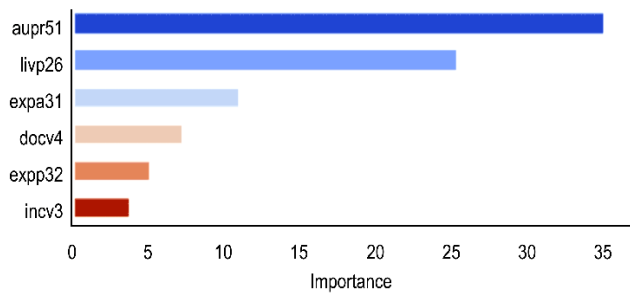


FIGURE 6. Feature importance Top 6 of the wholesale and retail industry.

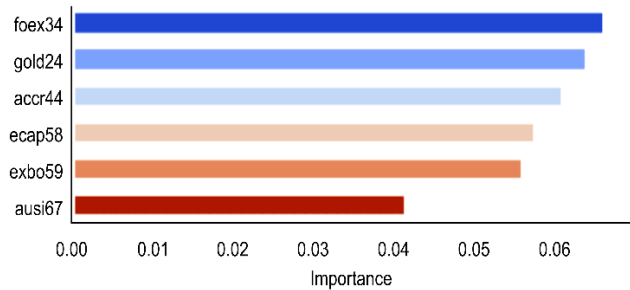


FIGURE 7. Feature importance Top 6 of the information and communication industry.

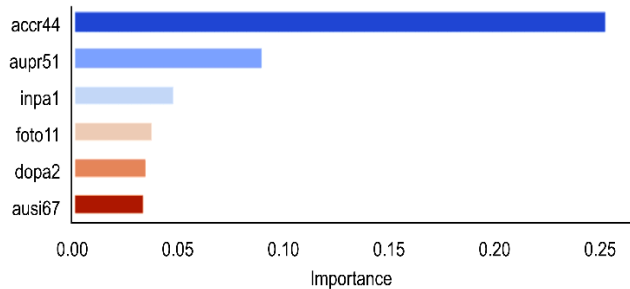


FIGURE 8. Feature importance Top 6 of the finance and insurance industry.

the number of automobiles produced shows the highest predictive power for wholesale and retail company sales.

Fig. 7 shows the importance of features in the information and communication industries. The descending order of feature importance is as follows: Foreign exchange reserves, gold, industrial accident rate, economically active population, external bonds, and automobile sales index. Among these, the foreign exchange reserves shows the highest predictive power for information and communication company sales.

Fig. 8 shows the feature importance of the finance and insurance industries. The descending order of feature importance is as follows: industrial accident rate, number of automobiles produced, number of international passengers, number of foreign tourists visiting Korea, number of domestic passengers, and automobile sales index. Among these, the industrial accident rate shows the highest predictive power for finance and insurance company sales.

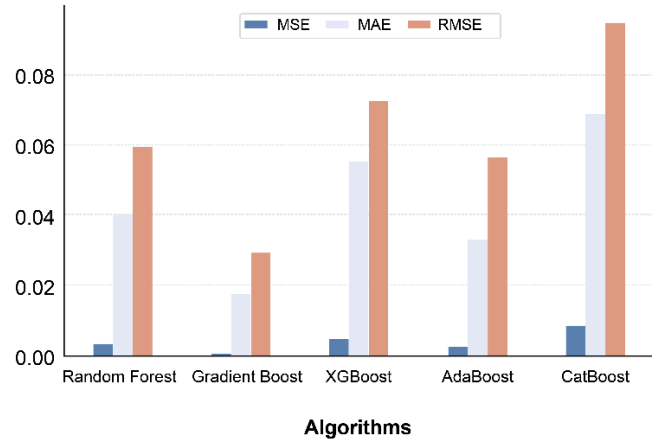


FIGURE 9. Comparison of the results of regression performance evaluation index by algorithm.

Table 10 summarizes the feature importance of the top six by industry.

Differences were found on checking the features by category in Table 3 regarding the feature importance by industry in Table 10. The manufacturing, finance, and insurance industries were found to be highly affected by the working conditions. In contrast, the wholesale, retail, and information and communication industries were highly affected by growth, and the construction industries were highly affected by raw materials.

### C. COMPARISON OF REGRESSION ANALYSIS RESULTS BY THE MACHINE LEARNING ALGORITHMS

The performance of the machine learning algorithms was compared to determine the most suitable algorithm for predicting the sales of KOSPI- and KOSDAQ-listed companies. Feature importance was compared to determine whether the variables that had the largest influence on the sales of listed companies and the variables that affected listed companies varied with the industry using the optimal machine learning algorithm. From the results, the following observations were made:

First, the results of the machine learning analysis of listed companies are shown in Fig. 9. Gradient Boost was the optimal machine learning algorithm according to the evaluation criteria of the regression performance evaluation indices MAE, MSE, and RMSE.

Second, the machine learning analysis results of listed companies by industry are shown in Figs. 10–12. According to the evaluation criteria of the regression performance evaluation indexes MAE, MSE, and RMSE, the optimal machine learning algorithm for each industry suggested Gradient Boost for manufacturing, construction, and information and communication, CatBoost for wholesale and retail, and Random Forest for finance and insurance.

Third, the results of checking the top six features with the highest importance among the listed companies showed that the variable with the greatest influence was the industrial

**TABLE 10. Feature importance Top 6 by industry.**

Industry	Ranking	Category	Feature	Description
Manufacturing	1	Working conditions	accr44	Industrial accident rate
	2	Employment	ecap58	Economically active population
	3	Production	whre61	Wholesale and retail index
	4	Growth	foex34	Foreign exchange reserves
	5	Working conditions	deto45	Occupational disaster death toll
	6	Custom import and export	expv33	Export volume index
Construction	1	Raw materials	gold24	Gold
	2	Price	livp26	Cost-of-living index
	3	Interest rate	bara69	Base rate
	4	Working conditions	accr44	Industrial accident rate
	5	Price	impp29	Import price index
	6	Working conditions	deto45	Occupational disaster death toll
Wholesale and retail	1	Growth	aupr51	Number of automobiles produced
	2	Price	livp26	Cost-of-living index
	3	Custom import and export	expa31	Export amount index
	4	Transportation infrastructure	docv4	Domestic cargo volume
	5	Price	expp32	Export price index
	6	Transportation infrastructure	incv3	International cargo volume
Information and communication	1	Growth	foex34	Foreign exchange reserves
	2	Raw materials	gold24	Gold
	3	Working conditions	accr44	Industrial accident rate
	4	Employment	ecap58	Economically active population
	5	Stability	exbo59	External bonds
	6	Consumption	ausi67	Automobile sales index
Finance and insurance	1	Working conditions	accr44	Industrial accident rate
	2	Growth	aupr51	Number of automobiles produced
	3	Transportation infrastructure	inpa1	Number of international passengers
	4	Leisure	foto11	Number of foreign tourists visiting Korea
	5	Transportation infrastructure	dopa2	Number of domestic passengers
	6	Consumption	ausi67	Automobile sales index

accident rate in the working conditions category, as shown in Table 8.

Fourth, the results of checking the top six features with the highest importance among the listed companies by industry showed that the variables with the highest influence varied by industry (Table 10). The most influential variables in the manufacturing, financial, and insurance industries were the industrial accident rate related to working conditions, in the construction industry was gold related to raw materials, in the wholesale and retail industries were the number of automobiles produced related to growth, and in information and

communication industries were foreign exchange reserves related to growth.

Table 11 summarizes the national statistical indicators that significantly affect the sales of listed companies and listed companies by industry based on the results of the machine learning regression analysis.

Consequently, Gradient Boost was confirmed to be the optimal machine learning algorithm for all industries when national statistical indicators predicted sales of KOSPI- and KOSDAQ-listed companies and that the optimal machine learning algorithm differed for each industry. In addition,



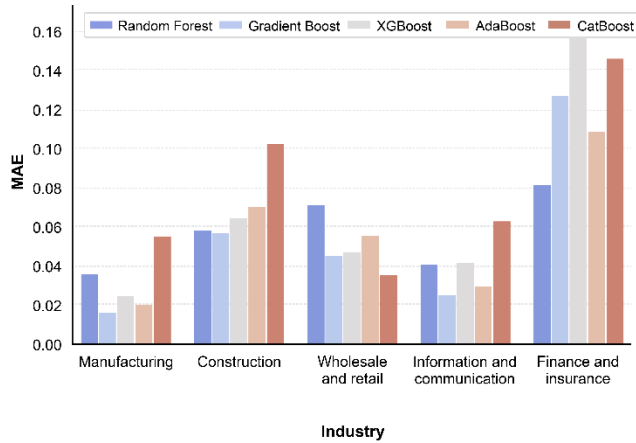


FIGURE 10. Comparison of MAE results by the industry for all algorithms.

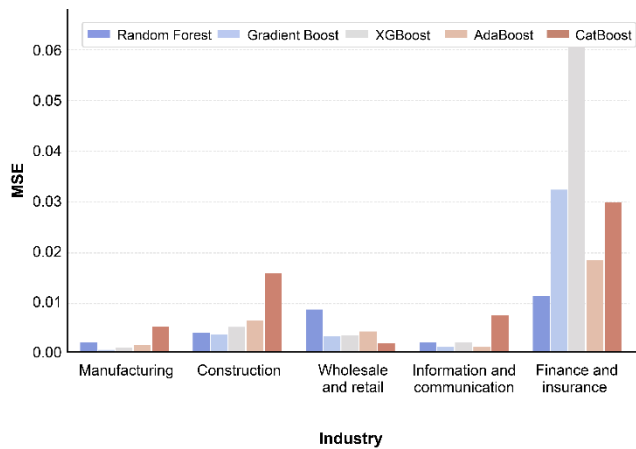


FIGURE 11. Comparison of MSE results by the industry for all algorithms.

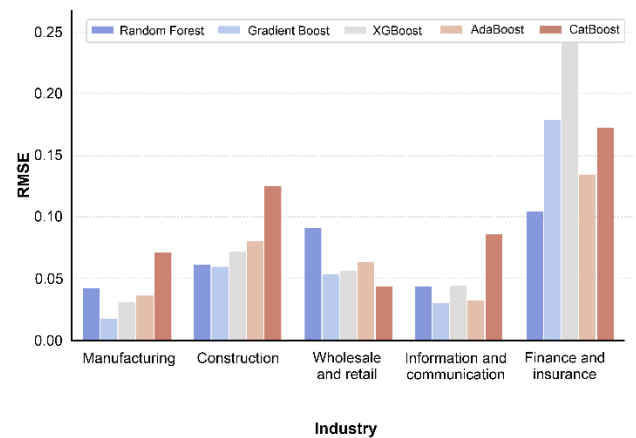


FIGURE 12. Comparison of RMSE results by the industry for all algorithms.

the most influential variables for listed companies were confirmed to vary with the industry.

These results are an empirical analysis of the limitations of related studies that did not identify the macroeconomic indicators affected by industry, and they were obtained using

TABLE 11. National statistical indicators that have a significant impact on the sales of listed companies and listed companies by industry.

Industry	Category	Feature	Description
Manufacturing	Working conditions	accr44	Industrial accident rate
Construction	Raw materials	gold24	Gold
Wholesale and retail	Growth	aupr51	Number of automobiles produced
Information and communication	Growth	foex34	Foreign exchange reserves
Finance and insurance	Working conditions	accr44	Industrial accident rate
The whole industry	Working conditions	accr44	Industrial accident rate

national statistical indicators, KOSPI- and KOSDAQ-listed companies' sales, thereby suggesting which national statistical indicators should be used as basic data to enhance management activities.

### V. CONCLUSION

This study aimed to investigate whether national statistical indices, created to understand social and economic conditions, affect company sales. Furthermore, it aimed to determine the optimal machine learning algorithm by exploring the variables that most affected the sales of KOSPI- and KOSDAQ-listed companies and by comparing the performance of several machine learning algorithms.

For this purpose, 58 national statistical indices provided by the Statistics Korea e-Nara Index and Bank of Korea Economic Statistical System and the annual sales of 1,470 KOSPI- and KOSDAQ-listed companies registered in the Korea Listed Companies Association with more than 20 years of history were used as analysis data.

Model learning and evaluation were performed using the machine learning algorithms RF, Gradient Boost, XGBoost, AdaBoost, and CatBoost. The importance of the features that influence the sales of KOSPI- and KOSDAQ-listed companies and the listed companies by industry were derived, and they were compared using a machine learning algorithm. The regression performance evaluation metrics of MAE, MSE, and RMSE were employed. After the analysis, the whole industry presented Gradient Boost as an optimal machine learning algorithm. By industry, manufacturing, construction, and information and communication industries presented Gradient Boost, wholesale and retail industries presented CatBoost, and finance and insurance industries presented Random Forest as the optimal machine learning algorithms.

In addition, upon examining feature importance to explore the variables that most affected the sales of listed companies

using the optimal machine learning algorithm, the industrial accident rate was found to have the greatest predictive power on the sales of listed companies among national statistical indicators. By industry, the variable with the highest predictive power was the industrial accident rate for the manufacturing, financial, and insurance industries, gold in the construction industry, the number of automobiles produced in the wholesale and retail industry, and foreign exchange reserves for the information and communication industry. This result confirmed that the national statistical indices that affected the sales of listed companies varied by industry.

In other words, it was confirmed that for a company to establish a management strategy, the national statistical indicators that have the most influence depending on the industry, which is a characteristic of the company, should be used as basic data.

This study determined the optimal machine learning algorithm for the variables and data characteristics of the national statistical indicators affecting the sales of KOSPI- and KOSDAQ-listed companies. However, the study has limitations, which are discussed below.

First, in terms of research subjects, KOSPI- and KOSDAQ-listed companies were compared and analyzed by industry. This is an analysis of companies with equity capital of more than 1.5 billion won among corporate size requirements and meeting KOSPI- and KOSDAQ-listing conditions. Companies are divided into large companies, medium-sized companies, small and medium-sized companies, and small companies according to the size of the company. KOSPI- and KOSDAQ-listed companies were only partially listed for large companies, medium-sized companies, and small and medium-sized companies, so small and medium-sized companies with an equity capital of less than 1.5 billion won were not included in the study. Owing to the limitations of the collected research subjects, national statistical indicators were analyzed according to the industry, which is a characteristic of a company, for KOSPI- and KOSDAQ-listed companies, but the effect of the national statistical indicators on the size of the company could not be analyzed. Therefore, the research subjects should be expanded according to the size of the company. The research results can be further generalized if national statistical indicators that affect corporate sales according to the size of the company are compared and analyzed using corporate sales data from corporate credit rating agencies.

Second, in terms of variables, 58 statistical indicators representing social and economic conditions were used as independent variables, and annual sales data of KOSPI- and KOSDAQ-listed companies were used as dependent variables. The Korea Listed Companies Association collects and manages sales data for KOSPI- and KOSDAQ-listed companies once a year according to the sales data collection cycle, so only annual sales data exist and monthly and quarterly sales data are not managed. Therefore, 58 statistical indicators providing annual indicators among 743 national

statistical indicators were used to derive predictive variables that affected sales of listed companies. However, there is a limit to generalizing the research results as predictive variables were derived only from 58 statistical indicators. Therefore, if the Korea Listed Companies Association collects the sales data of KOSPI- and KOSDAQ-listed companies from once a year to once a quarter, 58 statistical indicators will include quarterly statistical indicators to expand the target of independent variables. In addition, predictive variables that affect the sales of KOSPI- and KOSDAQ-listed companies will be expanded.

By expanding and examining the variables, and research objects, which are the limitations of this study, companies will be able to provide a considerable amount of data, which can be used to develop management strategies.

This study is the first to empirically analyze the national statistical indicators a company should use to establish management strategies, and it is expected to provide a good foundation for follow-up studies using the sales growth of companies and the national statistical indices.

## REFERENCES

- [1] L.-C. Zhang, "Proxy expenditure weights for consumer price index: Audit sampling inference for big-data statistics," *J. Roy. Stat. Soc. Ser. A, Statist. Soc.*, vol. 184, no. 2, pp. 571–588, Apr. 2021, doi: [10.1111/rssa.12632](https://doi.org/10.1111/rssa.12632).
- [2] M. Chen and X. Hu, "Linkage between consumer price index and purchasing power parity: Theoretic and empirical study," *J. Int. Trade Econ. Develop.*, vol. 27, no. 7, pp. 729–760, Oct. 2018, doi: [10.1080/09638199.2018.1430164](https://doi.org/10.1080/09638199.2018.1430164).
- [3] *e-Nara Index*. Accessed: Jun. 24, 2022. [Online]. Available: <https://www.index.go.kr/main.do>
- [4] Y. Sova and I. Lukianenko, "Theoretical and empirical analysis of the relationship between monetary policy and stock market indices," in *Proc. 10th Int. Conf. Adv. Comput. Inf. Technol. (ACIT)*, Sep. 2020, pp. 708–711, doi: [10.1109/ACIT49673.2020.9208926](https://doi.org/10.1109/ACIT49673.2020.9208926).
- [5] N. B. Fernandez and R. C. Li, "The influence of macroeconomic variables on philippine stock market indices: A structural equation model approach," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, Dec. 2020, pp. 1027–1031, doi: [10.1109/IEEM45057.2020.9309816](https://doi.org/10.1109/IEEM45057.2020.9309816).
- [6] M. Guven, B. Cetinguc, B. Guloglu, and F. Calisir, "The effects of daily growth in COVID-19 deaths, cases, and governments' response policies on stock markets of emerging economies," *Res. Int. Bus. Finance*, vol. 61, Oct. 2022, Art. no. 101659, doi: [10.1016/j.ribaf.2022.101659](https://doi.org/10.1016/j.ribaf.2022.101659).
- [7] O. K. Loang and Z. Ahmad, "Market overreaction, firm-specific information and macroeconomic variables in U.S. and Chinese markets during COVID-19," *J. Econ. Stud.*, vol. 49, no. 8, pp. 1548–1565, Oct. 2022, doi: [10.1108/JES-10-2021-0543](https://doi.org/10.1108/JES-10-2021-0543).
- [8] G. Liu, X. Fang, Y. Huang, and W. Zhao, "Identifying the role of consumer and producer price index announcements in stock index futures price changes," *Econ. Anal. Policy*, vol. 72, pp. 87–101, Dec. 2021, doi: [10.1016/j.eap.2021.07.009](https://doi.org/10.1016/j.eap.2021.07.009).
- [9] D. G. Singh, "Machine learning models in stock market prediction," *Int. J. Innov. Technol. Exploring Eng.*, vol. 11, no. 3, pp. 18–28, Jan. 2022, doi: [10.35940/ijitee.C9733.0111322](https://doi.org/10.35940/ijitee.C9733.0111322).
- [10] V. Bhardwaj, P. Bhavsar, and D. Patnaik, "Forecasting GDP per capita of OECD countries using machine learning and deep learning models," in *Proc. Interdisciplinary Res. Technol. Manag. (IRTM)*, Feb. 2022, pp. 1–6, doi: [10.1109/IRTM54583.2022.9791714](https://doi.org/10.1109/IRTM54583.2022.9791714).
- [11] A. Chatterjee, H. Bhowmick, and J. Sen, "Stock price prediction using time series, econometric, machine learning, and deep learning models," in *Proc. IEEE Mysore Sub Sect. Int. Conf.*, Oct. 2021, pp. 289–296, doi: [10.1109/MysuruCon52639.2021.9641610](https://doi.org/10.1109/MysuruCon52639.2021.9641610).
- [12] G. Lee, "Exploring the predictive variables of major economic-related statistical indicators on SME sales using machine learning: Focusing on small and medium-sized businesses in print-related shopping," *J. Korean Assoc. Comput. Educ.*, vol. 25, no. 3, pp. 79–89, 2022, doi: [10.32431/kace.2022.25.3.007](https://doi.org/10.32431/kace.2022.25.3.007).

- [13] G. Lee, "Exploring the predictive variables of government statistical indicators on retail sales using machine learning: Focusing on pharmacy," *J. Internet Comput. Serv.*, vol. 23, no. 3, pp. 125–135, 2022, doi: [10.7472/jksii.2022.23.3.125](https://doi.org/10.7472/jksii.2022.23.3.125).
- [14] L. Gasparėnienė, R. Remeikiene, A. Sosidko, and V. Vėbraitė, "Modelling of S&P 500 index price based on U.S. economic indicators: Machine learning approach," *Eng. Econ.*, vol. 32, no. 4, pp. 362–375, Oct. 2021, doi: [10.5755/J01.EE.32.4.27985](https://doi.org/10.5755/J01.EE.32.4.27985).
- [15] V. Bharat, M. Sharma, and A. Saxena, "Modelling the Nexus of macro-economic variables with WTI Crude Oil Price: A Machine Learning Approach," in *Proc. IEEE Region 10 Symp. (TENSYP)*, Jul. 2022, pp. 1–6, doi: [10.1109/TENSYP54529.2022.9864544](https://doi.org/10.1109/TENSYP54529.2022.9864544).
- [16] C. M. Kwon, *Python Machine Learning Complete Guide*. New York, NY, USA: Wikibook, 2020.
- [17] M. Oh, "On the need for data linkage in the health and welfare sectors," *Health Welf. Policy Forum*, vol. 9, pp. 17–28, Jan. 2015.
- [18] Y. Jeong, "A study on the innovation plan of the ICT statistical production system in response to changes in the survey environment (II) general report," Korea Inf. Soc. Develop. Inst., 2017, pp. 1–237.
- [19] K. An, "Developing a prediction model for firm innovation and performance using statistical matching and machine learning ensemble techniques," Ph.D. dissertation, Manag. Inf., Dongguk Univ., Seoul, South Korea, 2021.
- [20] J. Kwon and M. Johnson, "Meaningful healthcare security: Does meaningful-use attestation improve information security performance?" *MIS Quart.*, vol. 42, no. 4, pp. 1043–1067, 2018.
- [21] P. Wei, Z. Lu, and J. Song, "Variable importance analysis: A comprehensive review," *Rel. Eng. Syst. Saf.*, vol. 142, pp. 399–432, Oct. 2015, doi: [10.1016/j.res.2015.05.018](https://doi.org/10.1016/j.res.2015.05.018).



**GWANGSU LEE** received the Ph.D. degree in computer education from Sungkyunkwan University, South Korea, in 2011. From 2005 to 2014, he was a Computer Education Expert with the Korea Advancing Schools Foundation under the Ministry of Education, South Korea. From 2015 to 2020, he was an Adjunct Professor with the Computer Education Department, Sungkyunkwan University. Since 2015, he has been an Expert of big data with the Korea SMEs and Startups

Institute under the Ministry of SMEs and Startups, South Korea. His current research interests include time series analysis, data mining, and big data analysis.

...