한국빅데이터학회지
제8권 제1호, 2023, pp. 35-47 https://doi.org/10.36498/kbigdt.2023.8.1.35

A Big Data-Driven Business Data Analysis System: Applications of Artificial Intelligence Techniques in Problem Solving

Donggeun Kim[1] · Sangjin Kim[2] · Juyong Ko[1] · Jai Woo Lee[1†]

Department of Big Data Science, College of Public Policy, Korea University[1]
Department of National Statistics, College of Public Policy, Korea University[2]

요 약

It is crucial to develop effective and efficient big data analytics methods for problem-solving in the field of business in order to improve the performance of data analytics and reduce costs and risks in the analysis of customer data. In this study, a big data-driven data analysis system using artificial intelligence techniques is designed to increase the accuracy of big data analytics along with the rapid growth of the field of data science. We present a key direction for big data analysis systems through missing value imputation, outlier detection, feature extraction, utilization of explainable artificial intelligence techniques, and exploratory data analysis. Our objective is not only to develop big data analysis techniques with complex structures of business data but also to bridge the gap between the theoretical ideas in artificial intelligence methods and the analysis of real-world data in the field of business.

■ Keyword : Artificial Intelligence; Big Data; Statistics; Lasso; Business Insights; Value of Customer Information

2023년 05월 22일 접수; 2023년 06월 13일 수정본 접수; 2023년 06월 16일 게재 확정.

† 교신저자 (jaiwoolee@korea.ac.kr)

Ⅰ. Introduction

The association of the features in big data for business can occur depending on potential common exposures and relationships [1]. In the case of firms, the use of big data in the financial industry is not so active yet. However, it is highly expected that the utilization of big data analytics will increase in the field of business because abundant data has been stored. Financial big data analytics enables the development of a recommendation system which can rapidly and accurately recognize whether consumers need a specific product [2]. Based on customer data and recommendation systems, a credit rating model can also be developed and extended.

This paper estimates the network of features related to customers and examines the changes in connectivity between features with or without data preprocessing [3]. The network is estimated using machine learning techniques after statistical pre-processing of customer data. Network estimation of customer features identifies a connection between features regarding customer satisfaction and loyalty. The biggest advantage of using big data is the visualization and prediction of customer behavior based on generalized properties [4]. Big data analytics can help collaborators to understand individual customer behavior more than traditional data analytics, possibly leading one-to-one marketing [5]. Studies on consumer behavior have been conducted in terms of improving existing methodologies for application to large networks of features related to customers [6,7].

Analyzing big data with complex structures requires detailed methodologies to understand the interactions and flows among features of the data. If there exist missing values in the data, there may be various reasons for the missingness of the data. Providing insufficient or inappropriate answers to the questions in a survey or conducting repeated experiments with inconsistent results may result in missing values in the data [8].

In developing a network model of customer data, it is assumed that a complex network with sparsity can be built based on information from data. Analyzing big data with complex structures can cause low computational speed in real-world data analysis [9]. Since network estimation in the field of business deals with complex economic and social structures, it is necessary to devise an efficient and effective methodology so that high-dimensional data can be applied with appropriate computational speed [10]. In this aspect, we can consider various categories of methods in big data analytics. Given a dataset including complex associations among features, outlier detection or unsupervised learning can be considered. Each feature may have at least a few hundred samples assuming a specific probability distribution with suspected outliers [11]. Suspected outliers indicate extremely small or large values possibly existing in the distribution of data values representing each feature or specific data values far from the trend in data values of multiple features [12]. When the suspected outliers may not affect data analysis results, it may be better to remove them. While outlier detection can reduce the number of samples, it may be important to extract and examine strongly interacting features in the data. The Graphical Lasso method was introduced to derive complex financial networks of features in the data at a feasible computational speed [13]. The utilization of Graphical Lasso builds a set of

strongly interacting or non-interacting features in the data, enabling the focused investigation of strongly interacting features [14]. To obtain features in the data which strongly affect the outcome of customer loyalty, GLMNET is used to fit generalized linear models on data via penalized maximum likelihood [15].

As penalized regression is applied to analyze high-dimensional datasets, the impact of missing data imputation on the improvement of accuracy in estimating a target outcome is considered. It is expected that appropriate missing data imputation may be able to track the trend in each feature and supplement the missingness of the data, possibly avoiding the over-fitting problem [8].

In the big data era, data is rapidly expanding in various academic fields. Developing cutting-edge computational methods which analyze high-dimensional data is needed. Furthermore, it is significant to measure and reduce the gap between the theoretical ideas in computational methods and the interpretation of real-world data analysis in the field of business. To resolve these issues, artificial intelligence techniques including missing data imputation, outlier detection, feature selection and prediction are presented.

The paper is organized as follows: Section 2 describes the network estimation method and sample data of characteristics in the management data used in this study. Section 3 discusses the performance of statistical pre-processing to better understand the structure of data. Section 4 shows the performance of penalized regression to estimate the outcome. Section 5 covers the conclusions and future research.

Ⅱ. Method

2.1 Overview

In section 2, missing data imputation, outlier detection, feature selection, and explainable artificial intelligence technique are introduced.

2.2 Missing Data Imputation

When missing values occur in the given data, the appropriate imputation of missing values should be used. With the fixed number of imputed datasets, an imputation method is suggested. In this case, the predictive mean matching (PMM) imputation method using a mice package in R was used. Other imputation methods can be used from a list of the available imputation methods [16].

2.3 Outlier detection

Outlier detection used principal components analysis to put data points on a two-dimensional mapping for visualization. The suspected outliers are detected using various statistical methods, possibly with prior knowledge from data. [17]

2.4 Feature Selection: GLASSO

In order to obtain an optimized network model, a more rigorous correlation is estimated by applying the Graphical Lasso method based on the overall correlation between business features. When defining the relationship between characteristics as a correlation, it is difficult to extract only important characteristics because all characteristics appear to have more than a certain relationship. However, using Graphical Lasso makes it easy to extract only important properties [18].

Considering a two-way network with n samples and p features, there are strongly related features. Each of the p features consists of n data values. We consider the proposed regression model to define the interaction between features and minimize the negative log-likelihood with penalties to obtain the optimal coefficients for this model. The minimization method follows Equation (1).

$$\log(\det(\theta)) - \mathrm{tr}(s\theta) + \rho||\theta||_1 \quad (1)$$

where S is the empirical covariance matrix, θ is a nonnegative definite matrix, tr indicates the trace of a covariance matrix, and $||\theta||_1$ is L1 norm. In Equation (2), β is estimated by minimizing the objective function in order to get an optimal hyperparameter ρ.

$$min_{\beta}\{\frac{1}{2}||W\beta - b||^2 + \rho||\beta||_1\} \quad (2)$$

2.5 Explainable Artificial Intelligence Technique: GLMNET

To estimate the credit risk network, we use a method of finding the association between business features considering covariance. When there is a specific correlation with some subsets of network characteristics, it is important to define the potential interactions of the features [19]. Therefore, we consider all the factors associated with the change and trend of the reputation level for each characteristic and verify the network model that defines the interaction between the two characteristics by combination by selecting a combination of the characteristics. Suppose there exist observations $x_i \in R^p$ and the responses $y_i \in$ R, i = 1, 2, …, N. The objective function for the Gaussian family follows Equation (3).

$$min_{(\beta_0,\beta)\in R^{p+1}}\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda[(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1] \quad (3)$$

where λ is nonnegative and α ranges from 0 to 1. If α is 0, a ridge regression is implemented, if α is 1, a lasso regression is implemented, and else, an elastic net regression is implemented. β is estimated by minimizing the objective function in order to get an optimal hyperparameter λ.

Ⅲ. Performance of Statistical Pre-Processing

3.1 Data

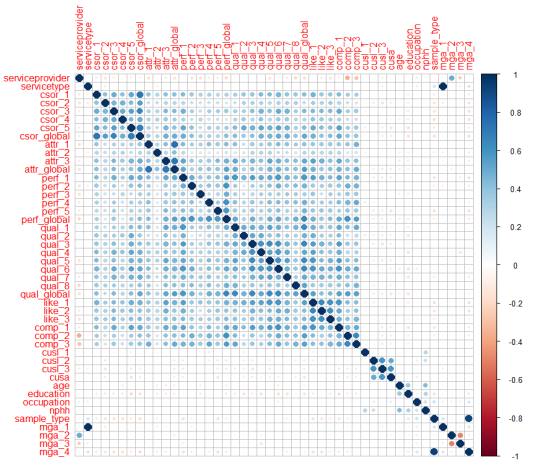
Five theoretically grounded candidate model configurations which consider customer satisfaction (CUSA) are presented [20]. Five alternate corporate reputation models are defined as follows: ATTR indicates Attractiveness,
CSOR indicates Corporate social responsibility, COMP indicates Competence,
CUSA indicates Customer satisfaction, CUSL indicates Customer loyalty, LIKE indicates Likeability, PERF indicates Performance, and QUAL indicates Quality.

The dataset consists of 46 features and 344 samples.

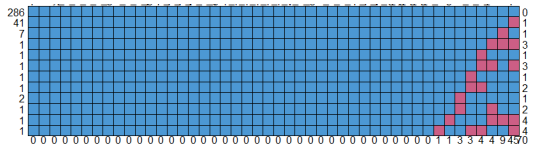
Dot points in <Figure 1> show the correlation between features in the data. A blue dot represents a positive correlation and a red dot represents a negative correlation. The darker the color, the higher the correlation.

3.2 Missing Data Imputation

As shown in <Figure 2>, there exist 70 missing



〈Figure 1〉 Correlation of Features in Data



〈Figure 2〉 Distribution of Missing Values

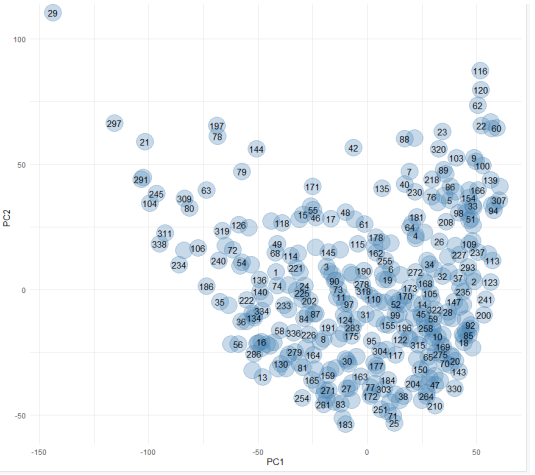
values in the original data, and these values are imputed using predictive mean matching methods. The missing values are imputed using values from the complete cases matched. The features including cusa, age, cusl_1, cusl_3, cusl_2, nphh, education, and occupation have missing values.

3.3 Outlier detection

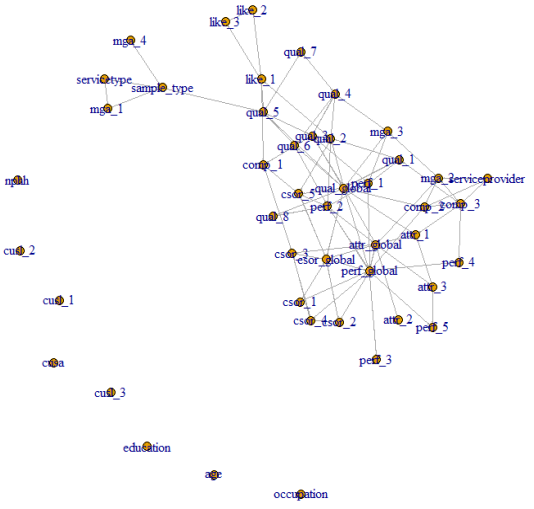
In <Figure 3> shown below, most values are clustered except for a few suspected outliers, such as a value of index 29. While removing the data value of index 29 does not greatly affect the data analysis results, this value was not removed to keep much information as possible.

3.4 Feature Extraction and Network Analysis

Regarding the feature extraction, network esti-



〈Figure 3〉 Detection of Suspected Outliers



〈Figure 4〉 Network Estimation before missing data imputation

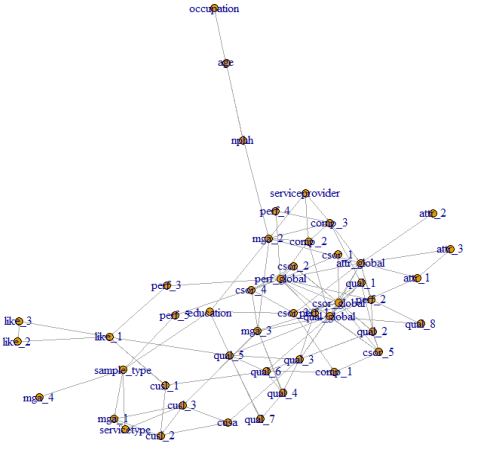
mations before and after missing data imputation using the predictive mean matching imputation method are implemented. As shown in <Figure 4>, features including cusa, age, cusl_1, cusl_3, cusl_2, nphh, education, and occupation form independent disconnected components. While all the other features form a large connected component, there do not exist associations among features outside this component in the network.

〈Table 1〉 Centrality analysis of the network in 〈Figure 4〉

	Degree	Closeness	Betweenness
serviceprovider	6	0.009	0
servicetype	4	0.007	0
csor_1	4	0.009	0
csor_2	6	0.009	1
csor_3	8	0.01	38.077
csor_4	6	0.01	6.348
csor_5	10	0.011	20.763
csor_global	14	0.012	69.509
attr_1	8	0.011	31.299
attr_2	2	0.009	0
attr_3	6	0.009	6.015
attr_global	18	0.013	174.936
perf_1	10	0.011	28.916
perf_2	14	0.013	70.774
perf_3	2	0.009	0
perf_4	4	0.009	6.286
perf_5	4	0.009	5.667
perf_global	24	0.013	321.573
qual_1	8	0.011	31.124
qual_2	14	0.012	114.943
qual_3	14	0.012	50.922
qual_4	10	0.01	23.405
qual_5	12	0.013	349.695
qual_6	12	0.012	39.999
qual_7	4	0.009	3.333
qual_8	4	0.01	0.333
qual_global	26	0.015	498.914
like_1	10	0.01	150.724
like_2	4	0.007	0
like_3	4	0.007	0
comp_1	10	0.01	50.786
comp_2	10	0.011	68.005
comp_3	12	0.01	29.786
sample_type	8	0.009	208
mga_1	4	0.007	0
mga_2	10	0.01	30.339
mga_3	8	0.011	30.53
mga_4	2	0.007	0

The features with high degree values greater than 20 in <Table 1> are qual_global and perf_global. Regarding the degree centrality, high values indicate that the features qual_global and perf_global may be central in the network structure. Since degree centrality represents how many connections each feature has, features with high degree values are connected to lots of other features at the center of the network. Closeness centrality shows how close a feature is to other features in the network. In this case, closeness centrality is obtained by computing the reciprocal of the sum of the shortest path length from the feature to every other feature in the network. The features with closeness greater than 0.01 are csor_5, csor_global, attr_1, attr_global, perf_1, perf_2, perf_global, qual_1, qual_2, qual_3, qual_5, qual_6, qual_global, comp_2, and mga_3. Betweenness centrality estimates how many shortest paths each feature is on. The features with betweenness greater than 100 are attr_global, perf_global, qual_2, qual_5, qual_global, like_1, and sample_type.

As shown in <Figure 5>, all features belong to a connected component and the features age and



〈Figure 5〉 Network Estimation after missing data imputation

〈Table 2〉 Centrality analysis of the network in 〈Figure 5〉

	Degree	Closeness	Betweenness
serviceprovider	6	0.007	0
servicetype	8	0.007	15.52
csor_1	4	0.008	0
csor_2	6	0.008	2.036
csor_3	10	0.009	76.895
csor_4	8	0.008	30.014
csor_5	8	0.008	15.445
csor_global	20	0.01	140.337
attr_1	6	0.008	14.067
attr_2	2	0.007	0
attr_3	4	0.007	0
attr_global	20	0.01	291.848
perf_1	10	0.009	38.882
perf_2	10	0.009	33.355
perf_3	4	0.008	67.12
perf_4	4	0.008	4.057
perf_5	4	0.008	72.983
perf_global	24	0.01	374.567
qual_1	10	0.008	24.557
qual_2	10	0.008	15.74
qual_3	16	0.009	159.751
qual_4	8	0.008	32.669
qual_5	12	0.01	242.922
qual_6	12	0.009	104.489
qual_7	4	0.007	3.515
qual_8	4	0.008	2.05
qual_global	24	0.011	391.984
like_1	10	0.008	197.719
like_2	4	0.006	0
like_3	4	0.006	0
comp_1	6	0.007	11.367
comp_2	10	0.009	68.621
comp_3	12	0.008	49.187
cusl_1	6	0.007	60.248
cusl_2	10	0.007	55.248
cusl_3	10	0.008	118.413
cusa	6	0.007	38.705
age	4	0.005	88
education	8	0.009	193.155
occupation	2	0.004	0
nphh	4	0.007	172
sample_type	10	0.008	167.178
mga_1	8	0.007	15.52
mga_2	14	0.009	348.51
mga_3	12	0.01	235.327
mga_4	2	0.006	0

occupation have an interaction. The features with high degree values greater than 20 in <Table 2> are qual_global and perf_global. The feature with closeness greater than 0.01 is qual_global only. The features with betweenness greater than 100 are csor_global, attr_global, perf_global, qual_3, qual_5, qual_6, qual_global, like_1, cusl_3, education, sample_type, mga_2, and mga_3. While the features with degree values greater than 20 are qual_global and perf_global regardless of missing data imputation, the closeness and betweenness values of features are affected by missing data imputation.

Ⅳ. Performance of Penalized Regression

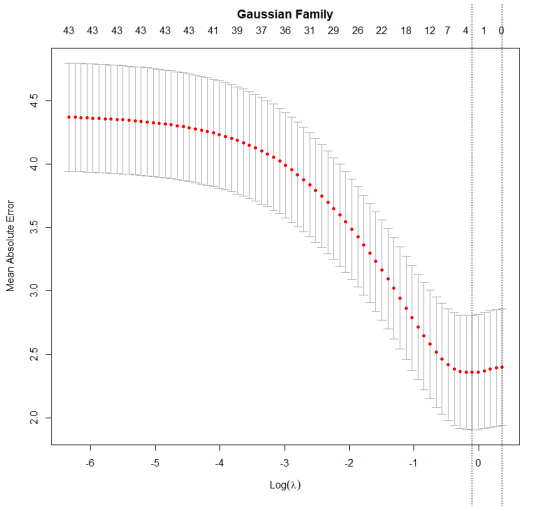
To identify the features which significantly affect the customer loyalty outcome, Lasso with 10-fold cross validation is utilized. Adding or removing a significant feature to the model possibly affects the properties of this feature only and might not affect other features in the model. In the real-world data, there may be interactions among multiple variables, and to implement Lasso, conditional independence between features should be considered.

Before PMM imputation method is applied, selecting significant features with lasso is not successful, not extracting features to build regression models.

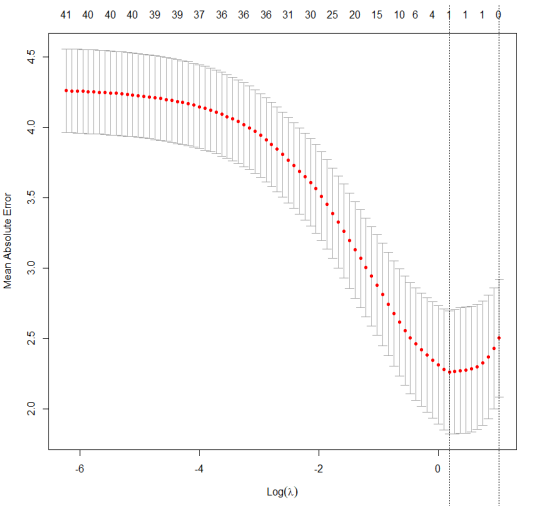
In <Figure 6>, when λ is 0.904, MAE is minimized and an optimal hyper-parameter is determined.

In <Figure 7>, when λ is 1.095, MAE is minimized and an optimal hyper-parameter is determined.

In <Figure 8>, when λ is 1.203, MAE is minimized and an optimal hyper-parameter is determined. The MAE results before applying the PMM im-



〈Figure 6〉 MAE results by log(λ) predicting cusl_1



〈Figure 8〉 MAE results by log(λ) predicting cusl_3

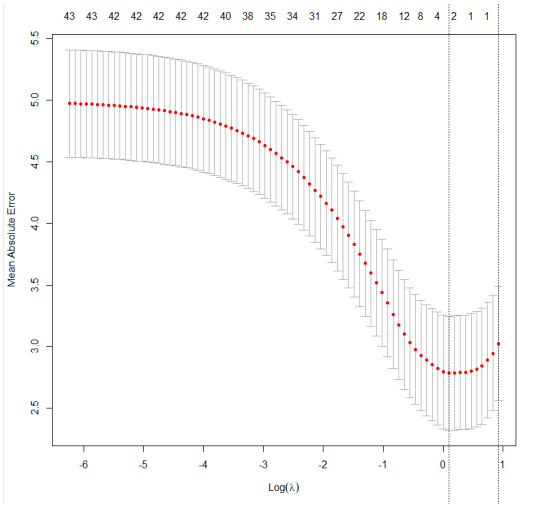
putation method on data show that finding optimal hyper-parameters can be better estimated when missing values are appropriately processed.

In <Table 3>, lasso regression identifies five variables such as perf_global, qual_6, like_1, like_2 and cusa which may positively affect the outcome, customer loyalty with an intercept value of 1.901.

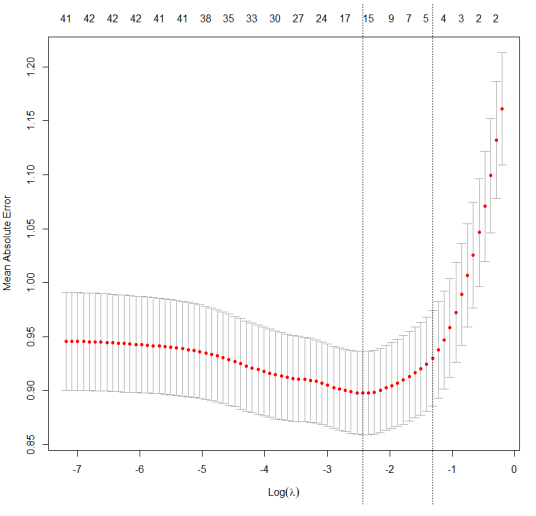
〈Table 3〉 Coefficients of features predicting cusl_1 after applying PMM imputation method

Intercept	perf_global	qual_6	like_1	like_2	cusa
2.247	0.007	0.072	0.178	0.034	0.279

In <Figure 9>, when λ is 0.088, MAE is minimized and an optimal hyper-parameter is determined.



〈Figure 7〉 MAE results by log(λ) predicting cusl_2

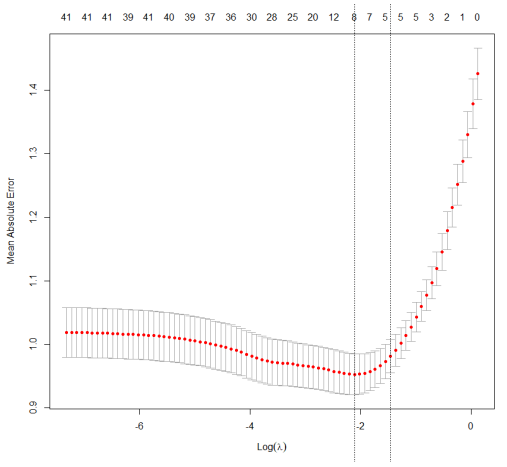


〈Figure 9〉 MAE results by log(λ) predicting cusl_1 after applying PMM imputation method

〈Table 4〉 Coefficients of features predicting cusl_2

Intercept	qual_3	like_1	like_2	like_3	cusa
0.875	0.059	0.013	0.130	0.063	0.588

In <Table 4>, lasso regression identifies five variables such as qual_3, like_1, like_2, like_3 and cusa which may positively affect the outcome, customer loyalty with an intercept value of 0.875.



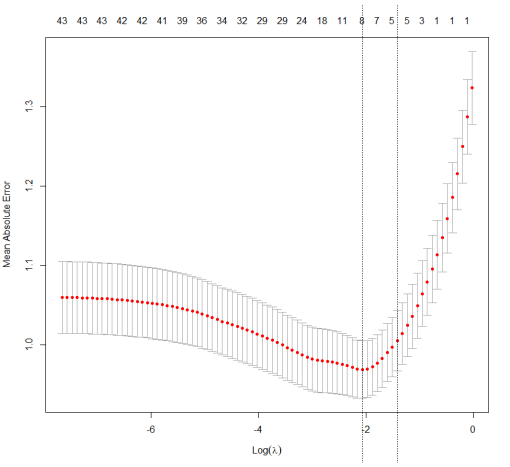
〈Figure 10〉 MAE results by log(λ) predicting cusl_2 after applying PMM imputation method

In <Figure 10>, when λ is 0.121, MAE is minimized and an optimal hyper-parameter is obtained.

〈Table 5〉 Coefficients of features predicting cusl_3

Intercept	qual_7	like_1	like_2	like_3	cusa
2.063	0.059	0.012	0.008	0.034	0.566

In <Table 5>, lasso regression identifies five variables such as qual_7, like_1, like_2, like_3 and cusa which may positively affect the outcome, customer loyalty with an intercept value of 2.195.



〈Figure 11〉 MAE results by log(λ) predicting cusl_3 after applying PMM imputation method

In <Figure 11>, when λ is 0.127, MAE is minimized and an optimal hyper-parameter is obtained. Mean absolute error results when zero imputation and imputation using PMM are applied. As shown in <Table 6>, using lasso regression on pre-processed data with PMM imputation method outperforms using lasso regression on pre-processed data with zero imputation method.

〈Table 6〉 MAE using zero and PMM imputation methods

	cusl_1	cusl_2	cusl_3
Zero	2.358	2.783	2.262
PMM	0.898	0.953	0.969

Ⅴ. Discussion & Conclusion

We have utilized data pre-processing methods, unsupervised and supervised learning techniques to elucidate the relations of features and appropriately predict an outcome, customer loyalty [21].

Based on the results from the prediction of customer loyalty, significant factors which affect customer loyalty may be performance, likeability, quality, and customer satisfaction.

Conducting a data preprocessing step that removes noise from the data and transforms the data into a useful format can fit the final model on all data values, establishing relationships between the relevant data features. It is required to meaningfully process raw data to get more organized data, and model accuracy [22]. Extracting significant features from data in the pre-processing process can improve the accuracy of the big data analysis, and the interpretation of the results may vary depending on which features are selected from the processed data. After the modeling process, the data preprocessing may be repeatedly conducted to better perform the model evaluation. Although with advanced analysis techniques, if data cleaning is not sufficiently implemented, less significant or potentially distorted analysis results can be generated. [23].

Network analysis has recently drawn much attention not only in physics, medicine, and biology but also in economics and business administration. Thanks to the accumulation of data with the development of information and communication technology via online services, the opportunity to use network data in practice has increased [24]. Internet networks, citation networks, or communication networks are structured in the form of networks consisting of nodes and edges indicating features and their interactions, respectively. Since networks have always provided a key environment to understand social and economic systems, network analysis on big data can provide core solutions to estimate dependencies between features in the network and to design the flow of the entire network [25]. Network analysis of big data is a method of analysis of structures, flows and processes regarding the relationship between features in the network. Unlike existing statistical research methods focusing on the properties of a feature, interactions among high-dimensional features are considered. Basic principles of network analysis from graph theory differ in the various fields of chemistry, biology, physics, information engineering, sociology, and humanities [26].

The big data analysis model should be designed using big data generated by the government or industry. It is necessary to handle customer data with both qualitative and quantitative approaches. More efficient customer management can be conducted by investigating customer's likeability about the product or customer satisfaction. The structures of data can be analyzed with a dataset with a low sample size by utilizing customer experience data of good quality [27]. Rather than a high-cost and time-consuming big data analysis method, an analysis model that observes and investigates a small number of significant customer data samples enable the understanding of the more complex customer purchase types [28]. Big data analysis examines customer experience and develops customer-centered products and services.

Future projects should develop ideas in the current study by building various conditions of simulated data. While a list of features in the data can be fixed, statistical experiments on data with different sample sizes can be conducted. Different combinations of missing data imputation methods, anomaly detection approaches, and network analysis techniques can be considered. Examining the performance of lasso regression models to predict

a specific continuous outcome can be tested in terms of accuracy and reproducibility. The validated methods can be utilized to analyze new outcomes from business data to obtain meaningful interpretations of data analysis results in various aspects.

사 사

This article is financially supported by the College of Public Policy at Korea University.

참 고 문 헌

[1] Jianqing Fan, Fang Han, Han Liu, Challenges of Big Data analysis, *National Science Review*, Volume 1, Issue 2, June 2014, Pages 293-314.

[2] Patrick Mikalef, John Krogstie, Ilias O. Pappas, Paul Pavlou, Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities, *Information & Management*, Volume 57, Issue 2, 2020, 103169.

[3] Fan C, Chen M, Wang X, Wang J and Huang B (2021) A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Front. Energy Res.* 9:652801.

[4] John Qi Dong, Chia-Han Yang, Business value of big data analytics: A systems-theoretic approach and empirical test, *Information & Management*, Volume 57, Issue 1, 2020, 103124.

[5] Sunil Erevelles, Nobuyuki Fukawa, Linda Swayne, Big Data consumer analytics and the transformation of marketing, *Journal of Business Research*, Volume 69, Issue 2, 2016, Pages 897-904.

[6] Rangaswamy, E., Nawaz, N. & Changzhuang, Z. The impact of digital technology on changing consumer behaviours with special reference to the home furnishing sector in Singapore. *Humanit Soc Sci Commun* **9**, 83 (2022).

[7] White, K., Habib, R., & Hardisty, D. J. (2019). How to SHIFT Consumer Behaviors to be More Sustainable: A Literature Review and Guiding Framework. *Journal of Marketing*, *83*(3), 22–49.

[8] Emmanuel, T., Maupong, T., Mpoeleng, D. *et al.* A survey on missing data in machine learning. *J Big Data* **8**, 140 (2021).

[9] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, Critical analysis of Big Data challenges and analytical methods, *Journal of Business Research*, Volume 70, 2017, Pages 263-286.

[10] Desamparados Blazquez, Josep Domenech, Big Data sources and methods for social and economic analyses, *Technological Forecasting and Social Change*, Volume 130, 2018, Pages 99-113.

[11] Goldstein M, Uchida S (2016) A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE* 11(4): e0152173.

[12] Panjei, E., Gruenwald, L., Leal, E. *et al.* A survey on outlier explanations. *The VLDB Journal* **31**, 977-1008 (2022).

[13] Kean Ming Tan, Daniela Witten, Ali Shojaie, The cluster graphical lasso for improved estimation of Gaussian graphical models, *Computational Statistics & Data Analysis*, Volume 85, 2015, Pages 23-36.

[14] Jain R, Xu W (2021) HDSI: High dimensional selection with interactions algorithm on feature

selection and testing. *PLOS ONE* 16(2): e0246159.

[15] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software, Articles* 33 (1): 1-22.

[16] Morris, T.P., White, I.R. & Royston, P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* **14**, 75 (2014).

[17] ur Rehman, A., Belhaouari, S.B. Unsupervised outlier detection in multidimensional data. *J Big Data* **8**, 80 (2021).

[18] Yusuke Hara, Junpei Suzuki, Masao Kuwahara, Network-wide traffic state estimation using a mixture Gaussian graphical model and graphical lasso, *Transportation Research Part C: Emerging Technologies*, Volume 86, 2018, Pages 622-638.

[19] Andrés Martínez, Claudia Schmuck, Sergiy Pereverzyev, Clemens Pirker, Markus Haltmeier, A machine learning framework for customer purchase prediction in the non-contractual setting, *European Journal of Operational Research*, Volume 281, Issue 3, 2020, Pages 588-596.

[20] Nicholas P. Danks, Pratyush N. Sharma, Marko Sarstedt, Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLS-SEM), Journal of Business Research, Volume 113, 2020, Pages 13-24.

[21] Mohammad Zoynul Abedin, Petr Hajek, Taimur Sharif, Md. Shahriare Satu, Md. Imran Khan, Modelling bank customer behaviour using feature engineering and classification techniques, *Research in International Business and Finance*, Volume 65, 2023, 101913.

[22] C.L. Philip Chen, Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Volume 275, 2014, Pages 314-347.

[23] Sarker, I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN COMPUT. SCI.* **2**, 377 (2021).

[24] Bickley, S.J., Chan, H.F. & Torgler, B. Artificial intelligence in the field of economics. *Scientometrics* **127**, 2055-2084 (2022).

[25] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, Giovanni Petri, Networks beyond pairwise interactions: Structure and dynamics, *Physics Reports*, Volume 874, 2020, Pages 1-92.

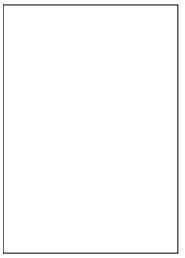
[26] Douglas A. Luke and Jenine K. Harris, Network Analysis in Public Health: History, Methods, and Applications, *Annual Review of Public Health* 2007 28:1, 69-93.

[27] Marko Sarstedt, Christian M. Ringle, Denis Iuklanov, Antecedents and consequences of corporate reputation: A dataset, Data in Brief, Volume 48, 2023, 109079.

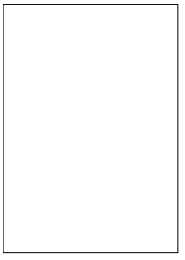
[28] Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. J Big Data. 2022;9(1):3

저 자 소 개

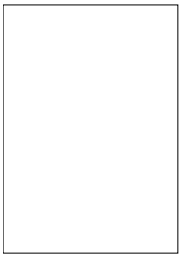
김 동 근(Donggeun Kim)
2018년 3월~현재: 고려대학교 공공정책대학 빅데이터전공
<관심분야>: 데이터사이언스, 인공지능, 컴퓨터비전



김 상 진(Sangjin Kim)
2018년 3월~현재: 고려대학교 공공정책대학 국가통계전공
<관심분야>: 데이터분석, 인공지능, 컴퓨터비전



고 주 용(Juyong Ko)
2019년 3월~현재: 고려대학교 공공정책대학 빅데이터전공
<관심분야>: 빅데이터, 데이터사이언스, 인공지능, 데이터 분석



이 재 우(Jai Woo Lee)
2012년 8월: 카네기멜론대학교 컴퓨터과학 (학사), 수학 (학사)
2016년 6월: 다트머스대학교 컴퓨터과학 (석사)
2019년 6월: 다트머스대학교 QuantitativeBiomedical Sciences (박사)
2022년 9월~현재: 고려대학교 공공정책대학 빅데이터사이언스학부 조교수
<관심분야> 데이터사이언스, 인공지능, 빅데이터