Information Systems Review
Vol. 24, No. 1 February 2022

http://dx.doi.org/10.14329/isr.2022.24.1.105

머신러닝 기반 부도예측모형에서 로컬영역의 도메인 지식 통합 규칙 기반 설명 방법

Domain Knowledge Incorporated Local Rule-based Explanation for ML-based Bankruptcy Prediction Model

조수현 (Soo Hyun Cho) 이화여자대학교 빅데이터분석학 박사과정
신경식 (Kyung-shik Shin) 이화여자대학교 경영대학 교수, 교신저자

요 약

신용리스크 관리에 해당하는 부도예측모형은 기업에 대한 신용평가라고도 볼 수 있으며 은행을 비롯한 금융기관의 신용평가모형의 기본 지식기반으로 새로운 인공지능 기술을 접목할 수 있는 유망한 분야로 손꼽히고 있다. 고도화된 모형의 실제 응용은 사용자의 수용도가 중요하나 부도예측모형의 경우, 금융전문가 혹은 고객에게 모형의 결과에 대한 설명이 요구되는 분야로 설명력이 없는 모형은 실제로 도입되고 사용자들에게 수용되기에는 어려움이 있다. 결국 모형의 결과에 대한 설명은 모형의 사용자에게 제공되는 것으로 사용자가 납득할 수 있는 설명을 제공하는 것이 모형에 대한 신뢰와 수용을 증진시킬 수 있다. 본 연구에서는 머신러닝 기반 모형에 설명력을 제고하는 방안으로 설명대상 인스턴스에 대하여 로컬영역에서의 설명을 제공하고자 한다. 이를 위해 설명대상의 로컬영역에 유전알고리즘(GA)을 이용하여 가상의 데이터포인트들을 생성한 후, 로컬 대리모델(surrogate model)로 연관규칙 알고리즘을 이용하여 설명대상에 대한 규칙기반 설명(rule-based explanation)을 생성한다. 해석 가능한 로컬 모델의 활용으로 설명을 제공하는 기존의 방법에서 더 나아가 본 연구는 부도예측 모형에 이용된 재무변수의 특성을 반영하여 연관규칙으로 도출된 설명에 도메인 지식을 통합한다. 이를 통해 사용자에게 제공되는 규칙의 현실적 가능성(feasibility)을 확보하고 제공되는 설명의 이해와 수용을 제고하고자 한다. 본 연구에서는 대표적인 블랙박스 모형인 인공신경망 기반 부도예측모형을 기반으로 최신의 규칙기반 설명 방법인 Anchor와 비교하였다. 제안하는 방법은 인공신경망 뿐만 아니라 다른 머신러닝 모형에도 적용 가능한 방법(model-agonistic method)이다.

키워드 : 부도예측모형, 로컬영역 설명력, 설명 가능한 인공지능(XAI)

† This work was supported by the National Research Foundation of Korea (NRF-2021R1A2C1012036).

Ⅰ. Introduction

Most of the recent studies focused on improving the performance of the financial prediction models using machine learning (ML) techniques. Machine learning techniques along with ensemble approach and deep learning were widely studied. Many studies focused on improving the performance of bankruptcy prediction and credit scoring models (Du Jardin, 2016; Feng *et al.*, 2018; He *et al.*, 2018; Marqués *et al.*, 2012; Moscatelli *et al.*, 2020). Compared to the number of studies focusing on the performance of the financial prediction models, only a small number of the studies focused on the interpretability of the ML-based models (Dastile *et al.*, 2020). No matter how accurate a model is, it is difficult to implement state-of-the-art machine learning models where high-stakes decisions are made. Those industries, such as finance, medicine and law, require and value explanation of the decisions. In highly regulated sectors like finance and medicine, models should balance both accuracy and explainability (Murdoch *et al.*, 2019). Furthermore, the General Data Protection Regulation (GDPR), which allows a user the right to explanation, went into effect in May 2018. At the same time, Basel II requires financial institutions to maintain a greater level of risk management. As a result, there is a growing demand for a model that is both accurate and interpretable. To overcome the shortcoming of ML techniques being “black-box” and to facilitate human understanding of the models, explainability now has become an important research topic, called Explainable Artificial Intelligence (XAI).

One of the well-known approaches to solve this problem is a rule-based explanation. The rule-based explanation can be achieved using rule-based learning. Rule-based learning refers to ML techniques that learns the patterns of the data by rules such as decision tree or random forests. The rule-based interpretable model was actively studied in the credit scoring model to provide explanations in a familiar rule format. Setiono and Liu (1996) extracted rules from a neural network-based model using symbolic representation of the neural network. Yi (2009) proposed a decision tree (C4.5) in conjunction with an approach called Simulating Annealing Algorithm (SAA) which performs global optimization for interpretable credit scoring. Hayashi (2016) proposed a recursive rule extraction algorithm with decision trees, called Re-Rx, to extract rules from ML-based credit scoring models. Soui *et al.* (2019) proposed a rule-based credit risk assessment model using multi-objective evolutionary algorithms. The author considered the generation of classification rules as an optimization problem. By using an evolutionary algorithm (EA), it aims to find the best combination of the customer characteristics and generate classification rules. Another paper proposed a two-stage rule extraction method based on a tree ensemble model for interpretable loan evaluation. Proposed tree ensemble model using two-stage rule extraction method.

In this study, we propose a local explanation generation method for the bankruptcy prediction model. The main contributions of the study include 1) a local explanation generation applied to a “black-box” bankruptcy prediction model using association rule mining algorithm as a local surrogate model and 2) generating feasible and informative explanations to the users by incorporating domain knowledge. In highly regulated sectors like finance and medicine, models should balance both accuracy and explainability (Murdoch *et al.*, 2019). With an increasing demand for the application of the XAI in the industry, this study applies a local explanation method to the ML-based bankruptcy prediction model. Also, most rule-based interpretability researches (Guidotti *et al.*, 2018; Hayashi, 2016; Rajapaksha *et al.*, 2020; Ribeiro and Guestrin, 2016; Ribeiro *et al.*, 2018; Soui *et al.*, 2019b) concentrated on the model's rule generation method and did not investigate if the generated rules were feasible in the real world. Some studies in counterfactual-based explanations focused on the feasibility of the generated explanation. Mahajan *et al.* (2019) proposed a method to offer feasible and actionable explanations by generating explanation that follows the underlying data distribution of the original data and Poyiadzi *et al.* (2020) adopted user labeling whether the generated explanation is feasible or not and trained an ML-based model to achieve feasibility of the generated explanation. The proposed model considers causal feasibility to improve the interpretability and comprehensibility of the explanation presented to the users by incorporating the causal feasibility of the financial variables used in the bankruptcy prediction model in the local explanation generation process. Here, the term "causal feasibility" refers to the possibility of certain states occurring in the real world given its current condition. To achieve explainability in ML models, the user of the system is always involved (Roscher *et al.*, 2020). If rules were to be used as an explanation of model output, it is important to have rules that make sense in the real world so the users with domain knowledge can understand and justify the result of the model prediction. Naturally, rules that seem not plausible in real life is difficult to justify the prediction made by the model. Especially, when there is an in-depth understanding of the domain by the system users. To tackle this issue, this paper generates association rules, both factual and counterfactual rules, and enhances the feasibility of generated rules by filtering rules using the causal relationship between financial variables and financial strength for the "black-box" bankruptcy prediction model.

Ⅱ. Related Studies

2.1 Explainability in Machine Learning

Technically speaking, there is no standard definition of the XAI (Adadi and Berrada, 2018). Also, explainability or interpretability are used interchangeably in the field (Adadi and Berrada, 2018; Carvalho *et al.*, 2019). XAI is more of a trend and movement towards AI transparency and trust issues. The goal of XAI can be clarified. Defense Advanced Research Projects Agency (DARPA) stated that XAI aims to provide models with more explainability while maintaining a high prediction accuracy (Gunning and Aha, 2019).

In general, XAI is imperative for users to understand and manage A. I. systems regardless of motivations. Main motivations or reasons for the need for XAI can be delivered into four categories according to Adadi and Berrada (2018). The first reason is to justify the machine learning model's decision. When it comes to the decision of a model, it is related to justification and reasons for a particular outcome rather than the logic of the model's inner mechanism to make a prediction. This can also ensure users that the model is dependable and fair. The second reason is to enhance the control over the prediction model itself by finding errors and correcting them. The third reason is to continuously improve the model. Since the users know why the model made such output, the model can be improved. The last reason comes from the need for discovery. Having explanations on the model output can be accumulated into knowledge and gain new insights about the model itself.

In recent years there has been an increasing number of research on the interpretability of ML techniques. The scope of explainability can be either global or local. Global explanation answers how the parts of the model affect predictions. This approach zooms into a model at a modular level and sees how it operates. Local explanation answers why did the model make a certain prediction for a certain instance. This approach zooms into a single instance and sees what the model prediction is and explains the prediction. In this paper, we focus on obtaining explanations in the format of rules for individual cases.

2.2 Local Approaches for Explanation

A number of studies suggested diverse methods generating an explanation of the ML-based at the local level. LIME (Ribeiro and Guestrin, 2016) presents an explanation with feature importance. For a given data point, LIME perturbs the feature values randomly and computes an approximate linear model to explain the prediction of the originally trained model. As an explanation, the coefficients of the features, representing the importance of the corresponding features, from the linear model are used. Similarly, SHAP (Lundberg and Lee, 2017) simulates the contribution each feature makes to the model, often explained as a collaborative multiplayer game setting, where the contribution of each player (i.e. feature) is measured by excluding the corresponding player from the game. Based on game theory, the method computes the average marginal contribution of each feature with a set of axiomatic properties that ensures fairness in the process.

LORE (Guidotti *et al.*, 2019), LoRMIkA (Rajapaksha *et al.*, 2020) and ANCHOR (Ribeiro and Guestrin, 2018) proposed a method to generate a rule-based explanation for each case to explain the model prediction. LORE uses a decision tree to clarify the local decision boundary and LoRMIkA uses a k-optimal class association rule mining method to mine rules for instance. LORE proposed a method to employ a decision tree on the synthetic neighborhood of the instance to be explained to derive rules that explain the reasons of the model prediction called decision rule and a set of counterfactual rules. LoRMIkA used an OPUS algorithm on the neighborhood of the instance to be explained to search k-optimal association rules to explain the model at the local model. The authors argued that the most predictive rules are not necessarily the best explanation and the interestingness of the rules should be considered as well along with predictiveness.

They adopted the OPUS algorithm as it captures infrequent higher-order associations which leads to interesting rules. To measure the interestingness of the rule, LoRMIkA used lift as an absolute difference from one. Anchor is a local rule-based explanation approach by generating a rule called anchors. Anchors are incrementally constructed. First, empty anchors (i.e. rules) are constructed and new candidate anchors are created extending anchors by one additional feature predicate. The final anchor is chosen with the highest estimated precision in the model. To find the anchors, the problem can be formulated using a multi-armed bandit problem or beam search approach. Such methods provided an explanation in a simple rule format and can be easily understood by the users. However, most of the studies neglected the feasibility of the generated explanation and a few studies considered the feasibility of the generated explanation with the user's help for feasible explanation generation. For example, Mothilal *et al.* (2020) applied post-hoc filtering of the explanation by the users and Mahajan *et al.* (2019) further trained the model to generate feasible explanation after labeling the generated explanation by the user. The proposed model suggests a feasible rule-based explanation for the bankruptcy prediction model by incorporating a causal relationship between financial variables and financial soundness that is consistent with the domain knowledge.

2.3 Explainability in Finance

Interpretability of machine learning models has been investigated in the field of finance for some time, ever since the deployment of opaque machine learning models. This comes from the fact that in the field of finance, even though the misclassification cost of the problem is high the domain requires users' understanding and without explanation on the model's output, it is very difficult to use the model in practice.

To generate interpretable models, simple intrinsic interpretable models were used in the past. Henley and Hand (1996) proposed a simple k-nearest neighbor for the credit scoring model. This method can easily provide why such prediction was made by the model using a simple algorithm searching for nearest neighbors. Similarly but more recently, Grath *et al.* (2018) proposed a counterfactual-based explanation to explain the result of loan applications. A counterfactual-based explanation is a type of local explanation generation method that uses a similar synthetic case yielding an opposite model prediction as an explanation. The study proposed two weighting strategies (i.e. feature importance and nearest neighbor) to generate more interpretable counterfactuals.

As the significance of the interpretability in ML-based models for financial prediction increases, Fair Isaac Corporation (FICO) launched the Explainable Machine Learning Challenge in 2018 in response to growing interest in XAI. The goal was to develop new research in the domain of algorithmic explainability with credit scoring data. Participants were challenged to create ML models that are both accurate and explainable, with aim of solving the credit scoring problem.

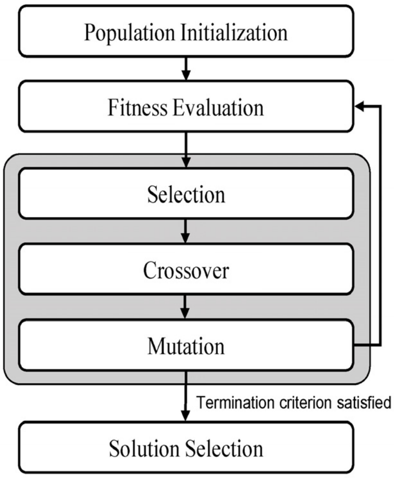
The winner, Dash *et al.* (2018), proposed Boolean Rules via Column Generation (BRCG), a global interpretable model for classification where Boolean rules in disjunctive normal form (DNF) or conjunctive normal form (CNF) are learned. Column generation is used to efficiently search through the number of candidate clauses without heuristic rule mining. BRCG dominates the accuracy-simplicity trade-off in half of the datasets tested, but even though it achieves good classification performance and explainability, methods like the RIPPER decision tree still obtain a better classification accuracy in many of the datasets, including HELOC. The authors state that one of the limitations includes performance variability as well as the reduced solution quality when implemented on large datasets. Gomez *et al.* (2020) used a Support Vector Machine (SVM) model with a linear kernel for classification model and proposed a method to find important features by systematical perturbation of the columns while holding others fixed. The method combines both local level explanations and global level model interpretations to visualize the logic behind the model's decisions to users with the most contributing features in a decision identified.

Ⅲ. Methodology

3.1 Genetic Algorithm

A GA is an evolutionary algorithm that is widely used for generating near-optimal solutions for search problems. As the name indicates, it is a general adaptive optimization search method derived from Darwin's evolution theory. To briefly explain the process of a GA, it starts with a set of populations comprising chromosomes, which each contain a certain number of genes. The GA conducts operations on the populations in an "evolutionary" way to search for the best chromosome. Each chromosome represents a solution and thus, a set of chromosomes, i.e., a population, indicates a set of possible solutions. The solutions are evaluated based on their fitness function, which is declared by a researcher in advance.

GA operations include selection, mutation, and crossover to generate generations of populations. This involves the algorithm using both the exploitation and exploration methods to search for potential solutions. The process of GA is illustrated in <Figure 1> and described below. In this study, GA was used to generate a local neighborhood of a data point to build a local explainer using the generated local neighborhood.



〈Figure 1〉 Genetic Algorithm (GA) Flowchart

Population Initialization. First, initial population is generated by randomly making a certain number of chromosomes comprise a certain number of genes, both of which are declared beforehand to begin. Here, each gene contains a feature value and the number of the genes in each chromosome is consistent with the number of features used in the model. In other words, chromosomes represent possible synthetic data points in this study. Large numbers of populations introduce more diversity by enlarging the search space; however, they tend to converge slowly. In contrast, small numbers of populations converge faster; however, the search space may not be adequate to obtain a near-optimal solution for the problem.

Fitness Evaluation. This step involves evaluating the chromosomes' fitness in terms of their probability of being selected for the next generation. The fitness function may vary depending on the problem to which the GA is applied. The higher the fitness evaluation is, the greater the chance of a solution (i.e., a chromosome) being selected for the next generation.

Selection. After all the chromosomes have been evaluated using a fitness function, a new generation is developed. This is when the GA operators (selection, crossover, and mutation) come into effect. In the selection process, chromosomes are selected based on their fitness using methods such as the roulette and tournament approaches. For instance, with the tournament method, several tournaments are held with randomly chosen chromosomes and the winner is selected based on the fitness value of each chromosome. Tournament selection is similar to tournament match as it includes several numbers of matches for final selection (i.e. winner). The size of the tournament indicates a number of chromosomes (i.e. players) for the tournament and several tournament matches are held afterward to select the chromosome with the highest fitness value. The tournament method was used in the experiment as the method can be easily implemented and adjusted (Sharma *et al.*, 2014). In this study, the tournament method with a size of four was used for selection.

Crossover. Next, the selected chromosomes are crossed over pairwise to generate new chromosomes. This process is performed 2/N times since it utilizes two chromosomes to generate new chromosomes (offspring). This process can also be performed using the k-point crossover method. The k-point crossover method involves stochastically pinning k points to parent chromosomes and the genes between points are swapped between the parent chromosomes. In this study, a two-point crossover was used.

Mutation. A mutation is an exploratory approach in a GA operation conducted to prevent trapping in local optima. With a given probability, the algorithm randomly mutates the genes in the chromosomes. While a high mutation probability may delay convergence, it may also prevent the fall to the local optimum. Until the GA's stopping criterion is satisfied, it goes returns to the fitness evaluation step and repeats the process. After satisfying the stopping criterion, the GA ends its operation.

3.2 Association Rule Mining

Association rule mining is an algorithm that discovers interesting frequent patterns in the dataset. Apriori (Agrawal *et al.*, 1993) is a popular algorithm to retrieve rules from large data, given some computational requirements. Support and confidence are the most known requirements applied to discover meaningful rules. The key elements of all Apriori algorithms are specified by the measures allowing to mine association rules which have support and confidence greater than user-defined thresholds. To briefly explain the process of association rule mining, let us consider I = {i_1, i_2,⋯, i_N} as a set of N unique items and let D be the database of transactions where each transaction T can be an item or set of items, subset of I. Each transaction is associated with a unique identifier. Let X and Y be the items or sets of items. Hence, an association rule is of the form: X⇒Y, where X⊆I, Y⊆I and X∩Y=∅. In the following sections, we present terminology and equations commonly associated with association rule mining.

$$Support(X) = \frac{|instance \in Dataset,\ with\ X|}{total\ \#\ of\ instances\ in\ D} \qquad (1)$$

$$Support(X \Rightarrow Y) = \frac{|instance \in Dataset,\ with\ X\ and\ Y|}{total\ \#\ of\ instances\ in\ D} \qquad (2)$$

$$Lift(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)} \qquad (3)$$

$$Confidence(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \qquad (4)$$

To evaluate the rules, support, lift and confidence are often considered. Support is the probability of X occurring in a transaction set D. For only X part (LHS) of the rule, support can be calculated using Eq.(1) and for both X and Y, it can be calculated using Eq. (2). If support is too low, it indicates that the itemset does not occur frequently so the rule cannot draw important information. Lift measures the occurrences of X and Y given that X and Y are independent using Eq. (3). Generally, a lift value above one is considered to be predictive and has a meaningful association. Confidence of a rule is the conditional probability that the subsequent Y is true given the predecessor X as shown in Eq. (4). Confidence shows how predictive the rule is as it is the probability of having a consequent Y given the antecedent X. The value is ranged between [0, 1] and the closer the value is to one the more predictive the rule is. The predictability of a rule can be measured using the confidence of the rule and the interestingness of a rule can be measured by the lift or the leverage. Furthermore, when the value of the confidence is high, the rule is said to be a predictive rule and when the value of the lift or the leverage is high, the rule is said to be an interesting rule.

Ⅳ. Proposed Model

4.1 Proposed Model Background and Overview

The proposed method aims to provide local rule-based explanations by generating local neighborhood of an instance and mining association rules. The proposed model considers the domain-specific causal (directional) relationship of the financial input variables to generate feasible rules for the users. The explanation generated for users should make sense to the users for them to understand the provided explanation. One of the major goals of the XAI is to justify the model decision to the users (Adadi and Berrada, 2018; Lipton, 2016). At the same time, the user is always involved to achieve explainability in the ML models (Roscher *et al.*, 2020). As mentioned in Guidotti *et al.* (2019), explanations should be as close as possible to the language of reasoning, which is formal logic and if a user can understand a simple format of logic, it is easy to construct narratives understood by users. For example, if a credit scoring model made a decision to decline one's loan and provided an explanation saying the loan would have been granted "if income is lower than current income", it hardly makes sense to the people and is not likely to trust the model's decision as it contradicts the domain knowledge. By incorporating the causal relationship of the financial input variables into the process of explanation generation process for each instance, rules that are plausible in the real world can be offered to the users to explain the model decision.

The intuition behind the proposed model, like other local approaches, is that the decision boundary for the black box can be arbitrarily complex over the whole data space, but in the local neighborhood of an instance, there is a high chance that the decision boundary is clear and simple to be captured by the local model (i.e. association rule mining). Association rule mining is essentially a local model as the mining algorithm considers only certain features and only certain values of these features. This means that only a subspace of the feature space is considered. In this paper, we adopt an association rule mining algorithm as a local explainer to find both interpretable and predictive rules by controlling the criteria of the rules. Also, we constrain the consequent to the label of the data to generated class association rules. Novak *et al.* (2009) discuss the differences between interpretability and predictability, by showing that the most predictive rules and the rules that explain best on a given dataset will be usually different. Using the example of a C4.5 decision tree for a predictive algorithm, they illustrate that redundant rules will be ignored, while in descriptive algorithms, redundant rules should be considered. On the other hand, highly predictive rules may result from false correlations in the training data, if they represent only a small number of examples. Such rules will be filtered out by an adequate descriptive algorithm accordingly, while a predictive algorithm may be forced to take such rules into account for the sake of completeness of the predictions.

Furthermore, one of the benefits of using an association rule mining algorithm is that the rules identified by the algorithm can be diverse yet easily filtered to serve the user's needs unlike other rule-based explainer such as decision trees or random forests. In addition, association rule mining offers different types of rules that can contribute to the user's understanding. The types of rules from the rule mining algorithm can be presented in <Table 1>. Factual rules are the rules that supports the current instance whereas counterfactual rules are the rules that contradict the current instance in terms of both antecedent and consequent. The antecedent of the factual rules can present important features that led to the current prediction of the model whereas the antecedent of the counterfactual rules can present important features that may contribute to changing the model prediction. Complementary factual rules are rules that support the current model decision and demonstrate potentially important features related to the current prediction of the model. Likewise, complementary counter-

〈Table 1〉 Types of Rules Generated from Association Rule Mining as an Explainer

Rule Type	Antecedent (LHS)	Consequent (RHS)	Description
Factual Rules	True	True	- Rules supporting the current instance - Indicates important feature(s) that led to the current preiction from the model
Complementary Factual Rules	False	True	- Rules supporting a current prediction of the model - Indicates potentially important feature(s) for the instance related to the current prediction - i.e.) Red flag (risk) variables for bankrupt instances and green flag (safe) variables for non-bankrupt instances
Counterfactual Rules	False	False	- Rules contradicting the current instance - Indicates important feature(s) to obtain a opposite prediction from the model
Complementary Counterfactual Rules	True	False	- Rules contradicting a current prediction of the model - Indicates potentially important feature(s) for the instance related to the opposite prediction - i.e.) Green flag (safe) variables for bankrupt instances and red flag (risk) variables for non-bankrupt instances

factual rules are rules that contradict the current model prediction for an instance, yet it may inform important features related to obtaining an opposite prediction from the model. For example, in the case of complementary factual rules, the antecedents of the rules can reveal potential red flag features (for bankrupt instances) or green flag features (for non-bankrupt instances) of the model. The proposed model can provide the user with more diverse and informative rule-based explanations by adopting an association rule mining algorithm as a local explainer for the ML-based bankruptcy prediction model. Additionally, rules against the causal feasibility are excluded to enhance the understanding of the generated explanation for better comprehensibility. Other rule evaluation metrics such as support, confidence and lift are considered to filter the rules provided to the user.

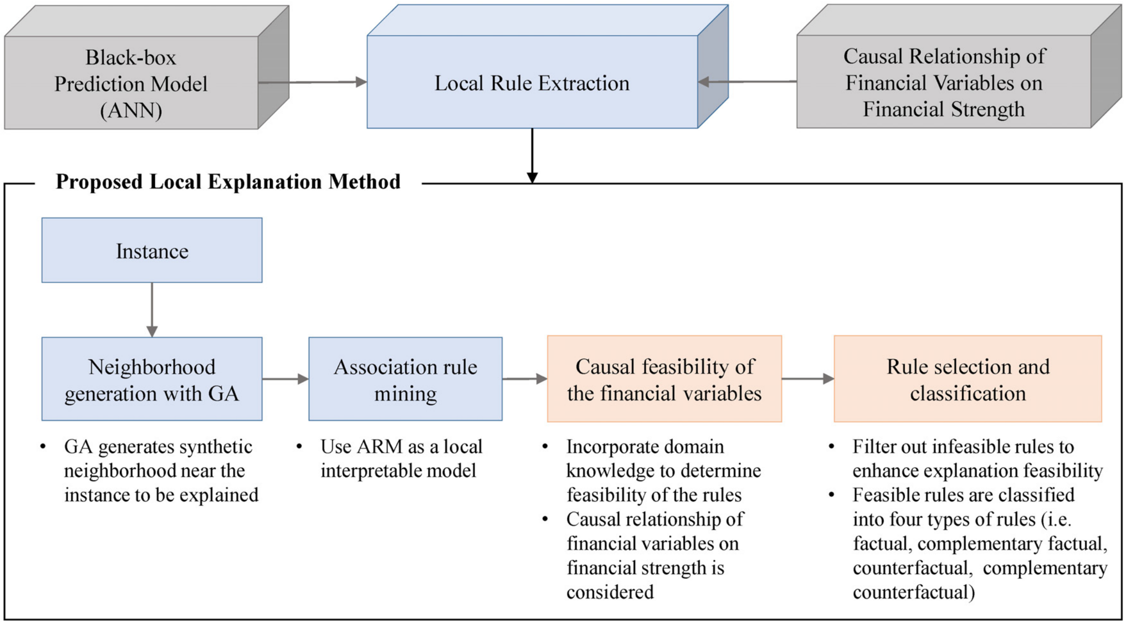
4.2 Proposed Model Process

The proposed model is a post-hoc explanation generation method, incorporating domain knowledge to the rule-based explanation to increase the feasibility of the rules and enhance the interpretability of the generated explanation. <Figure 2> shows the overall flowchart of the proposed method. As shown in the top part of the figure, a bankruptcy prediction model is trained at first, then the proposed local rule extraction method is implemented as a post-hoc measure to generate explanation. The bottom part of the figure illustrates more detailed process of the proposed explanation generation method. The proposed model incorporates domain knowledge to the rule-based explanation to increase the feasibility of the rules and enhance the interpretability of the generated explanation. In the proposed method, local neighborhood is generated with GA as LORE (Guidotti *et al.*, 2018) utilized GA to create local neighbors near the instance to be explained. To generate a local neighborhood, consisting of close synthetic instances, both positive and negative instances of the sample close to the original data distribution are created using GA. Each chromosome in the population represents synthetic neighbors of the instance to be explained and genes in the chromosome represent the feature values of the synthetic instance. Fitness function to obtain such local neighborhood has two terms; loss and distance as presented in Eq. (5).

$$\arg\min loss(x,\hat{x})\lambda_1 + dist(x,\hat{x})\lambda_2 \tag{5}$$

When generating a neighborhood with the same label as x, the $loss(x, \hat{x})$ yields zero only if the label of the x and $\hat{x}$ is the same, and otherwise zero. Likewise, when generating a neighborhood with the opposite label of x, the $loss(x, \hat{x})$ yields zero only if the label of the x and $\hat{x}$ is different and zero otherwise. $dist(x, \hat{x})$ is a distance term to measure the distance between the original instance x and new synthetic neighbor instance $\hat{x}$. For a distance metric, Euclidean distance is used. Also, λ is adopted as a trade-off parameter between the two terms presented in the fitness function.

After generating a local neighborhood, class association rule mining will be used as a local model. To implement a class association rule algorithm, feature values are discretized into three. All features are discretized to represent the directional change (i.e. upward or downward or within the range of ± 10% of the original instance’s feature value) compared to the original feature value. This can also offer more comprehensibility by providing rule antecedent in range, not in a specific fixed number. Also, it can represent upward or downward changes compared to the original instance to be explained as well. Lastly, rule selection and rule classification phase are implemented. Rule selection is carried out by filtering infeasible rules based on the causal (directional) relationship of the features and in the rule classification phase, rules are classified into four types of rules to provide more informative and feasible rule-based explanations.


Black-box
Prediction Model
(ANN)
Local Rule Extraction
Causal Relationship of
Financial Variables on
Financial Strength
Proposed Local Explanation Method
Instance
Neighborhood generation with GA
Association rule mining
Causal feasibility of the financial variables
Rule selection and classification
• GA generates synthetic neighborhood near the instance to be explained
• Use ARM as a local interpretable model
• Incorporate domain knowledge to determine feasibility of the rules
• Causal relationship of financial variables on financial strength is considered
• Filter out infeasible rules to enhance explanation feasibility
• Feasible rules are classified into four types of rules (i.e. factual, complementary factual, counterfactual, complementary counterfactual)

〈Figure 2〉 Proposed Model Flowchart

Ⅴ. Experiment

5.1 Experiment Setting

We tested the proposed model with data containing 4838 of bankrupt cases and non-bankrupt cases. The dataset is balanced and the observations consist of small-to-medium-sized manufacturing firms in Korea for five years between 2003-2007. After eliminating features with missing values, and financial variables with redundant meanings, 27 financial ratios were left and the final input features of 16 were chosen using the stepwise feature selection method. <Table 2> shows the selected input features and corresponding causal feasibility related to financial strength. To identify commonly acknowledgeable causal direction for the financial variables, we referred to the studies focusing on the analysis of corporate financial statements and ratios (Bank of Korea, 2020; Davidson, 2019; Helfert, 2001). For instance, the inventory turnover variable has an upward positive relationship with financial strength so if the change in inventory turnover is upward (i.e. increased) it is more probable to say the company is not suffering financially than to say it is. For the bankruptcy prediction model, an artificial neural network was used and grid-search method was employed to find the parameters of ANN as shown in <Table 3>. The architecture of the trained model is shown in <Table 3> and the model had two hidden layers with 16 and 12 neurons in each layer. For classification, sigmoid function was used in the output layer of the model. The trained model presented an accuracy of 80.51% for the training set and 79.82% for the test set as presented in <Table 4>.

〈Table 2〉 Selected Variables and Causal Direction

#	Variable No.	Variable Name	Category	Causal Direction
1	v11	Inventory turnover	activity	up
2	v110	Working capital requirement (KRW)	activity	down
3	v17	Non-current asset turnover	activity	up
4	v19	Working capital cycle (days)	activity	down
5	v26	Owner’s capital growth	growth	up
6	v37	Gross value-added to machinary (KRW)	productivity	up
7	v39	Gross value added per capita (KRW)	productivity	up
8	v411	Operating income to sales	profitability	up
9	v415	Retained earnings to total asset	profitability	up
10	v418	CGS to sales	profitability	down
11	v423	Debt service coverage ratio (DSCR)	profitability	up
12	v47	Net income on shareholder’s equity	profitability	up
13	v51	Cash ratio	stability	up
14	v513	Non-current assets to shareholders’ equity and non-current liabilities	stability	down
15	v515	Current liability ratio	stability	down
16	v56	Financial cost to sales	stability	down

〈Table 3〉 Model Architecture for Bankruptcy Prediction Model

Parameters	Details
Number of layers	2
Number of neurons	16, 12
Activation function	relu
Optimizer	Adam
Alpha	0.0001
Epoch	max. 500

〈Table 4〉 Bankruptcy Prediction Model Performance

Metric	Training set	Test set
Accuracy	0.8051	0.7982
Precision	0.8109	0.8105
Recall	0.8032	0.7768
F-1 score	0.8051	0.7938

5.2 Experiment and Result

After training the prediction model, a rule-based local explanation generation algorithm was applied. The local explanation was generated on the test set. For each instance to be explained, 2000 neighborhood instances consisting of 1000 bankrupt and 1000 non-bankrupt class labels were generated using GA. The fitness function in the GA uses loss and distance of the original instance and generated population to acquire neighborhood for an instance to be explained. The population of the GA for each label generation was set to 1000. In this study, we used the population of the final generation as a local neighborhood. Therefore, the number of total populations equals the number of neighborhoods with the aimed label, 1000 for each label. To initialize the population, each gene in each chromosome was set to change from the value of the original instance to be explained with the probability of 0.5 to begin the search with a population close to the original instance. We used the tournament method and two-point crossover method for selection and crossover. GA operation was set to terminate when it reached its maximum generation of 20. <Table 5> demonstrates the detailed parameter setting for the GA used in the experiment. Using the generated synthetic local neighborhood, a class association rule mining algorithm was applied to find rules that are consistent with the domain knowledge to explain the model decision.

〈Table 5〉 Parameter Setting for GA

Parameters	Details
Fitness function	arg min $loss(x, \hat{x})\lambda_1 + dist(x, \hat{x})\lambda_2$
Population	1,000
Max. generation	20
Selection	Tournament
Mutation rate	0.7
Crossover rate	0.2 (two-point crossover)

$\lambda_1 = 1,\ \lambda_2 = 0.6$

To compare the proposed method, the state-of-the-art rule-based local explanation method Anchor (Ribeiro *et al.*, 2018) was used as a benchmark model. Anchor generates a single rule as an explanation for each instance which is called an anchor. The goal of the method is to generate rules with high precision in the prediction model that supports the current prediction of the model.

<Table 6> demonstrates an example of the rules extracted from the proposed method and Anchor.
An instance used for this case was predicted to be ‘bankrupt’ by the global prediction model. Below explains how the rules can be interpreted. The rule implies which features can potentially affect the model prediction. Therefore, a rule can be used by the users to explain or recommend to a customer which features should be changed to have or avoid a certain model prediction. Counterfactual rules from the proposed

method can be a useful tool to explain and suggest alternative changes to obtain a desirable prediction from the model.

A rule extracted from Anchor can be interpreted as: IF the company had working capital cycle (v19) longer than 67.06 days and shorter or equal to 91.25 days AND working capital requirement (v110) larger than 381,573,260.50 (KRW) and smaller or equal to 557,351,471.90 (KRW) AND inventory turnover (v11) larger than 22.56 and smaller or equal to 65.79 THAN the model prediction made is ‘bankrupt’. Anchor delivers ranges of feature values that will result in the same model prediction holding other features constant.

A factual rule from the proposed method can be interpreted as: IF cash ratio (v51) is 4.08 (current value or within the range of ± 10%) AND financial cost to sales (v56) 1.05 (current value or within the range of ± 10%) THEN the model prediction is “bankrupt”. This type of rule can explain important features that potentially affected the model to yield current prediction.

A counterfactual rule from the proposed method can be interpreted as: IF Non-current asset turnover (v17) increased more than 10% from the current value of 2.83 AND current liability ratio (v515) decreased more than 10% from the current value of 28.99 THEN the model will be “non-bankrupt”. Counterfactual rules from the model imply important features in the model that can yield different model prediction. From this rule, we can assume Non-current asset turnover (v17) and current liability ratio (v515) were considered important in the model’s decision for this case.

〈Table 7〉 Experiment Result

Measure	Anchor	Proposed Method
Rule length (Number of items in LHS)	4.75	**2.20**
Confidence	0.0429	**0.9368**
Coverage (LHS support)	**0.9722**	0.5351
Lift	0.0429	**1.8800**

The performance of the local explainers are presented in <Table 7>. The best values for each evaluation metric are in bold. We considered rule length, confidence (i.e. precision), coverage and lift to evaluate the performance of the local explainers. We used rule length as one of the evaluation metrics as a proxy to measure the quality of the rule and lift to evaluate how strong the association is between LHS and RHS, following LoRMIkA (Rajapaksha *et al.*, 2020). At the same time, commonly used evaluation metric for

〈Table 6〉 Generated Rule-Based Explanation Example

Method	Rule	Confidence	Coverage	Lift

Anchor	IF 67.06 < v19 <= 91.25 AND
	381573260.50 < v110 <= 557351471.90 AND
	22.56 < v11 <= 65.79
	THEN ‘Bankrupt’
		0.0972	0.9755	0.10

Proposed Method - Factual Rule	IF v51 = 4.08 AND
	v56 = 1.05
	THEN ‘Bankrupt’
		0.9968	0.5211	1.84

Proposed Method - Counterfactual Rule	IF v17 > 2.83 AND
	v515 < 28.99
	THEN ‘Non-Bankrupt’
		0.9200	0.4648	1.91

rule-based explanation generations methods (Guidotti *et al.*, 2018; Rajapaksha *et al.*, 2020; Ribeiro *et al.*, 2018), confidence and coverage were used. A detailed explanation on the evaluation is presented in the following.

Rule length is the average number of items in the antecedent of the generated rules. A smaller number of items in LHS yields simpler rules to understand. LoRMIkA, which utilized OPUS search-based association rules, also used an average number of features used in the rule-based explanation as a proxy to measure the interpretability of the explanation by the simplicity of the generated explanation. LORE also used tree depth, which is the length of the rule derived from decision tree to measure the complexity of the generated explanation. As the result shows, the proposed method has 2.20 items (i.e. features) in the antecedent of the rules on average. The anchor had the longest average rule length of 4.75.

Coverage measures the fraction of the neighborhood samples that satisfy the antecedent of the rule (i.e. support of LHS). The result showed that Anchor had the highest coverage across other methods whereas the proposed method presented the lowest coverage. We believe that GA used for neighborhood generation yielded relatively diverse neighbors close to the instance to be explained.

Confidence shows the fraction of the neighborhood instances that satisfy both antecedent and consequent of the rule out of instances with the antecedent condition, so it shows how predictive the rule is. The proposed method showed 0.9668 of confidence, which is the highest among the other methods. This shows that the proposed method was able to generate predictive rules compared to Anchor.

Lift can be calculated by dividing the confidence by the unconditional probability of the consequent. It measures how much more often the rule antecedent and consequent occur together if they were statistically independent. A lift greater than one indicates that the occurrence of the rule antecedent and consequent is more significant than it would be if the two were independent. In this experiment, the lift value of one is the minimum required lift for the class association rules. The result showed that the proposed method was able to generate rules with 1.88 of lift on average. It should be noted that the lift of Anchor is the same as its confidence as the method only uses neighbors that shares the same prediction as an instance to be explained. In other words, if an instance to be explained has the model prediction of ‘bankrupt’ then Anchor uses synthetic neighbors with ‘bankrupt’ predictions.

Ⅵ. Conclusion

There have been many approaches to shed a light on the obscurity of ML-based models with interpretable rules to explain the global model yet there the feasibility of the rules generated and domain knowledge were neglected. However, we believe that it is important to acquire the feasibility of the generated explanation for “black-box” models to enhance the interpretability of the explanation to the users. The proposed model uses the directional causal feasibility of the financial variables to incorporate the domain knowledge. We believe that an explanation that is acceptable and understandable in the real world by the users convey more interpretability. In that sense, this study conducted a practical application of the local explanation generation method on the bankruptcy prediction model and proposed a method to incorporate financial domain knowledge.

The proposed method incorporates domain knowledge to the rule-based explanation generation process to provide users with an explanation that suits the industry knowledge. To extract rules from the prediction model, the proposed method used association rule mining as a local explainer. Unlike other rule-based models, association rule mining can offer more informative rules by offering various types of rules such as factual, complementary factual, counterfactual and complementary counterfactual rules. The experiment showed that the proposed method can provide feasible rules consistent with the domain knowledge and the quality of the rules, confidence and rule length outperformed Anchor.

To provide an explanation on model prediction to human users, it is important to offer an explanation that "makes sense" in the real world. Feasible rules that do not contradict the domain knowledge offer users with justifiable rule-based explanations of a "black-box" neural network-based bankruptcy prediction model compared to those that contradict the domain knowledge. Focusing on this point, this paper proposed a local rule-based explanation for the bankruptcy prediction model by incorporating financial domain knowledge. The proposed method can offer an understandable explanation and enhance the interpretability of the ML-based model to users using the bankruptcy prediction model in the industry by providing feasible rules. Also, the proposed method is a model-agonistic approach that can be applied to models with other techniques. Association rules can be easily implemented with a simple mechanism and easy to control the quality conditions of the rules although it can be computationally costly. However, this study did not use the optimized set of parameters in GA operation for neighborhood generation. Future research can focus on improving the computational cost of the model by adopting other rule extraction algorithms or integrating rule mining algorithm with the rule selection phase. In addition, more complex relationships between the input variables can be considered to refine the feasibility of the explanation.

References

[1] Adadi, A. and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)", *IEEE Access*, Vol.6, 2018, pp. 52138-60.

[2] Agrawal, R., T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases", *ACM SIGMOD Record*, Vol.22, No.2, 1993, pp. 207-16.

[3] Bank of Korea, Financial Statement Analysis for 2019, Bank of Korea, 2020.

[4] Carvalho, D. V., E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics", *Electronics*, Vol.8, No.8, 2019, pp. 1-34.

[5] Dash, S., O. Günlük, and D. Wei, "Boolean decision rules via column generation", *Advances in Neural Information Processing System 2018-Decem*, 2018, pp. 4655-4665.

[6] Dastile, X., T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey", *Applied Soft Computing Journal*, Vol.91, No.106263, 2020, pp. 1-21.

[7] Davidson, W., *Financial Statement Analysis Basis For Management Advice, Association of International Certified Professional Accountants*, Inc., NC, USA, 2019.

[8] Feng, X., Z. Xiao, B. Zhong, J. Qiu, and Y. Dong, "Dynamic ensemble classification for credit scoring using soft probability", *Applied Soft Computing Journal*, Vol.65, 2018, pp. 139-51.

[9] Gomez, O., S. Holter, J. Yuan, and E. Bertini, "ViCE: Visual counterfactual explanations for machine learning models", *25th International Conference on Intelligent User Interfaces*, 2020.

[10] Grath, R. M., L. Costabello, C. Le Van, P.

Sweeney, F. Kamiab, Z. Shen, and F. Lécué, "Interpretable credit application predictions with counterfactual explanations", *NIPS 2018 Workshop on Challenges and Opportunities for AI In Financial Services: The Impact of Fairness, Explainability*, Accuracy, and Privacy, 2018.

[11] Guidotti, R., A. Monreale, S. Ruggieri, F. Giannotti, D. Pedreschi, and F. Turini, "Factual and counterfactual explanations for black box decision making", *IEEE Intelligent Systems*, 2019, pp. 14-23.

[12] Guidotti, R., A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems", arXiv preprint arXiv:1805.10820, 2018.

[13] Gunning, D. and D. W. Aha, "DARPA's explainable artificial intelligence program", *AI Magazine*, Vol.40, No.2, 2019, pp. 44-58.

[14] Hayashi, Y., "Application of a rule extraction algorithm family based on the Re-RX Algorithm to financial credit risk assessment from a pareto optimal perspective", *Operations Research Perspectives*, Vol.3, 2016, pp. 32-42.

[15] He, H., W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios", *Expert Systems with Applications*, Vol.98, 2018, pp. 105-17.

[16] Helfert, E. A., "Assessment in Business Performance", in Helfert, E. A. (1st ed.), *Financial Analysis Tools and Techniques: A Guide for Managers*, McGraw-Hill, NY, New York, 2001, pp. 95-160.

[17] Henley, W. E. and D. J. Hand, "A k-nearest-neighbour classifier for assessing consumer credit risk", *The Statistician*, Vol.45, No.1, 1996, pp. 77-95.

[18] Jardin, P. D., "A two-stage classification technique for bankruptcy prediction", *European Journal of Operational Research*, Vol.254, No.1, 2016, pp. 236-52.

[19] Lipton, Z. C., "The Mythos of Model Interpretability," *2016 ICML Workshop on Human Interpretability in MachineLearning*, 2016

[20] Mahajan, D., C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers", *33rd Conference on Neural Information Processing Systems*, 2019.

[21] Marqués, A. I., V. García, and J. S. Sánchez, "Two-level classifier ensembles for credit risk assessment", *Expert Systems with Applications*, Vol.39, No.12, pp. 10916-22, 2012.

[22] Moscatelli, M., F. Parlapiano, S. Narizzano, and G. Viggiano, "Corporate default forecasting with machine learning", *Expert Systems with Applications*, Vol.161, No.113567, 2020, pp. 1-12.

[23] Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning", *Proceedings of the National Academy of Sciences of the United States of America*, Vol.116, 2019, pp. 22071-80.

[24] Novak, P. K., N. Lavrač, and G. I. Webb, "Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining", *Journal of Machine Learning Research*, Vol.10, 2009, pp. 377-403.

[25] Poyiadzi, R., K. Sokol, R. Santos-rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations", *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

[26] Rajapaksha, D., C. Bergmeir, and W. Buntine, "LoRMIkA: Local Rule-Based Model Interpretability with k-Optimal Associations", *Information Sciences*, Vol.540, 2020, pp. 221-41.

[27] Ribeiro, M. T. and C. Guestrin, "'Why should

I trust you?' Explaining the predictions of any classifier", KDD 2016, 2016.

[28] Ribeiro, M. T., S. Singh, and C. Guestrin, "Anchors: High-Precision model-agnostic explanations", *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 1527-35.

[29] Roscher, R., B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries", *IEEE Access*, Vol.8, 2020, pp. 42200-216.

[30] Setiono, R. and H. Liu, "Symbolic representation of neural networks", *Computer*, 1996, pp. 71-77.

[31] Sharma, P., A. Wadhwa, and K. Komal, "Analysis of selection schemes for solving an optimization problem in genetic algorithm", *International Journal of Computer Applications*, Vol.93, No.11, 2014, pp. 1-3.

[32] Soui, M., I. Gasmi, S. Smiti, and K. Ghédira, "Rule-base d credit risk assessment model using multi-objective evolutionary algorithms", *Expert Systems With Applications*, Vol.126, 2019, pp. 144-57.

[33] Yi, J., "Credit scoring model based on the decision tree and the simulated annealing algorithm", *2009 WRI World Congress on Computer Science and Information Engineering*, 2009, pp. 18-22.

Information Systems Review
Volume 24 Number 1
February 2022

Domain Knowledge Incorporated Local Rule-based Explanation for ML-based Bankruptcy Prediction Model

Soo Hyun Cho* · Kyung-shik Shin**

Abstract

Thanks to the remarkable success of Artificial Intelligence (A.I.) techniques, a new possibility for its application on the real-world problem has begun. One of the prominent applications is the bankruptcy prediction model as it is often used as a basic knowledge base for credit scoring models in the financial industry. As a result, there has been extensive research on how to improve the prediction accuracy of the model. However, despite its impressive performance, it is difficult to implement machine learning (ML)-based models due to its intrinsic trait of obscurity, especially when the field requires or values an explanation about the result obtained by the model. The financial domain is one of the areas where explanation matters to stakeholders such as domain experts and customers. In this paper, we propose a novel approach to incorporate financial domain knowledge into local rule generation to provide explanations for the bankruptcy prediction model at instance level. The result shows the proposed method successfully selects and classifies the extracted rules based on the feasibility and information they convey to the users.

Keywords: Terms - Bankruptcy Prediction, Local Explanation, XAI

* Ph.D. Candidate, Department of Big Data Analytics, Ewha Womans University
** Corresponding Author, Professor, School of Business, Ewha Womans University

◐ 저 자 소 개 ◑

조 수 현 (soohyuncho7117@gmail.com)

중앙대학교 경영학부에서 학사학위를 취득하였으며, 현재 이화여자대학교 대학원 빅데이터분석학 박사과정 중이다. 주요 관심분야는 금융분야 및 산업에서의 빅데이터 및 인공지능 응용, 설명 가능한 인공지능(XAI) 등이다.

신 경 식 (ksshin@ewha.ac.kr)

연세대학교에서 경영학 학사와 George Washington University에서 MBA 학위를 받았으며, KAIST에서 경영공학 박사학위를 취득하였다. 현재 이화여자대학교 경영학부 교수로 재직 중이며, Expert Systems with Applications, Journal of Management Information Systems, Journal of Information Science, Journal of Computer Information Systems, Applied Soft Computing 등에 논문을 발표하였다.

논문접수일 : 2021년 11월 23일　　게재확정일 : 2022년 01월 24일
1차 수정일 : 2022년 01월 10일