http://www.jsebs.org
pISSN 2288-3908
eISSN 2765-3846

The Journal of Society for e-Business Studies
Vol.26, No.3, August 2021, pp.97-117
https://doi.org/10.7838/jsebs.2021.26.3.097

설명 가능한 정기예금 가입 여부 예측을 위한 앙상블 학습 기반 분류 모델들의 비교 분석

A Comparative Analysis of Ensemble Learning-Based Classification Models for Explainable Term Deposit Subscription Forecasting

신지안(Zian Shin)*, 문지훈(Jihoon Moon)**, 노승민(Seungmin Rho)***

초 록

정기예금 가입 여부 예측은 은행의 대표적인 금융 마케팅 중 하나로, 은행은 다양한 고객 정보를 활용하여 예측 모델을 구성할 수 있다. 정기예금 가입 여부의 분류 정확도를 향상하기 위해, 많은 연구에서 기계학습 기법들을 이용하여 분류 모델들을 개발하였다. 하지만, 이러한 모델들이 만족스러운 성능을 보일지라도 모델의 의사결정 과정에 대한 근거가 적절하게 설명되지 않는다면 산업에서 활용하기가 쉽지 않다. 이러한 문제점을 해결하기 위해, 본 논문은 설명 가능한 정기예금 가입 여부 예측 기법을 제안한다. 먼저, 테이블 형식에서 우수한 성능을 도출하는 의사결정 나무 기반 앙상블 학습 기법인 랜덤 포레스트, GBM, XGBoost, LightGBM을 이용하여 분류 모델들을 개발하고, 10겹 교차검증을 통해 모델들의 분류 성능을 심층 분석한다. 다음으로, 가장 우수한 성능을 도출하는 모델에 설명 가능한 인공지능 기법인 SHAP을 적용하여 고객 정보의 영향도와 의사결정 과정 등을 해석할 수 있는 근거를 제공한다. 제안한 기법의 실용성과 타당성을 입증하기 위해, Kaggle에서 제공한 은행 마케팅 데이터 셋을 대상으로 모의실험을 진행하였으며, 데이터 셋 구성에 따라 GBM과 LightGBM 모델에 SHAP을 각기 적용하여 설명 가능한 정기예금 가입 여부를 위한 분석 및 시각화를 수행하였다.

ABSTRACT

Predicting term deposit subscriptions is one of representative financial marketing in banks, and banks can build a prediction model using various customer information. In order to improve the classification accuracy for term deposit subscriptions, many studies have been conducted based on machine learning techniques. However, even if these models can achieve satisfactory performance, utilizing them is not an easy task in the industry when their decision-making process is not adequately explained. To address this issue, this paper

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2018-0-01799) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019R1F1A1060668).

* First Author, Master Student, Department of Security Convergence Science, Chung-Ang University (lanta4825@cau.ac.kr)
** Co-Author, Postdoctoral Researcher, Chung-Ang University(johnny89@cau.ac.kr)
*** Corresponding Author, Associate Professor, Department of Industrial Security, Chung-Ang University (smrho@cau.ac.kr)

Received: 2021-07-26, Review completed: 2021-08-10, Accepted: 2021-08-18

proposes an explainable scheme for term deposit subscription forecasting. For this, we first construct several classification models using decision tree-based ensemble learning methods, which yield excellent performance in tabular data, such as random forest, gradient boosting machine (GBM), extreme gradient boosting (XGB), and light gradient boosting machine (LightGBM). We then analyze their classification performance in depth through 10-fold cross-validation. After that, we provide the rationale for interpreting the influence of customer information and the decision-making process by applying Shapley additive explanation (SHAP), an explainable artificial intelligence technique, to the best classification model. To verify the practicality and validity of our scheme, experiments were conducted with the bank marketing dataset provided by Kaggle; we applied the SHAP to the GBM and LightGBM models, respectively, according to different dataset configurations and then performed their analysis and visualization for explainable term deposit subscriptions.

키워드 : 금융 마케팅, 정기예금 가입 예측, 설명 가능한 인공지능, 앙상블 학습, 배깅, 부스팅
Financial Marketing, Term Deposit Subscription Forecasting, Explainable Artificial Intelligence, Ensemble Learning, Bagging, Boosting

1. 서 론

정기예금이란 고객(예금자)이 금융기관과 사전에 일정 기간을 정하여 원금과 이자를 수취하기 위한 목적을 갖는 금융상품이다. 금융기관은 지급준비제도에 따라 예금자의 예금액 지급준비율을 제외한 나머지 금액을 이용하여 정기적금보다 많은 금액을 투자 또는 대출 사업을 위한 자본금으로 운용할 수 있으므로, 높은 경제적 이득을 기대할 수 있다. 그리하여, 최근 금융기관은 기존 고객들을 대상으로 정기예금 유치를 위해 다양한 마케팅 전략을 수립하고 있으며[9], 효과적인 금융 마케팅을 위해서는 기존의 고객 정보를 바탕으로 예측 모델을 구성한 뒤에 잠재적인 고객을 파악하는 것이 매우 중요하다[32].

여기서 정기예금 가입 여부는 범주예측 또는 분류(Classification)에 속하는 지도학습(Supervised Learning)의 일종이며, 나이, 직업, 자산 등 다양한 고객 정보를 고려하여 인공지능(Artificial Intelligence, AI) 기반의 분류 기법들이 보고되고 있다. Moro[22]는 SVM(Support Vector Machine), 의사결정 나무(Decision Tree, DT), 인공 신경망(Artificial Neural Network, ANN)을 이용하여 3가지 분류 모델들을 개발하였다. 실험 결과, ANN은 SVM, DT보다 우수한 분류 성능을 도출하였다.

국내에서도 Moro[22]와 동일한 데이터 셋을 이용하여 분류 모델을 개발한 사례가 보고되었다. Lee and Hwang[14]은 K-Means Clustering을 통해 나이, 잔금 등 수치형 변수를 군집화하여 학습 데이터를 구성하였으며, ANN, Naive Bayes, AdaBoost, 랜덤 포레스트(Random Forest, RF)를 이용하여 분류 모델을 개발하였다. 3겹 교차검증을 통한 실험 결과, ANN은 다른 분류 모델들보다 더 정확한 분류 성능을 도출하였다. Park et al.[27]은 기존 ANN 모델에서 은닉층의 수를 늘린 심층 신경망(Deep Neural Network, DNN) 모델을 구성하였으며, 기존 ANN 모델인 Lee and Hwang[14]보다 더 정확한 분류 성능을

도출하였다.

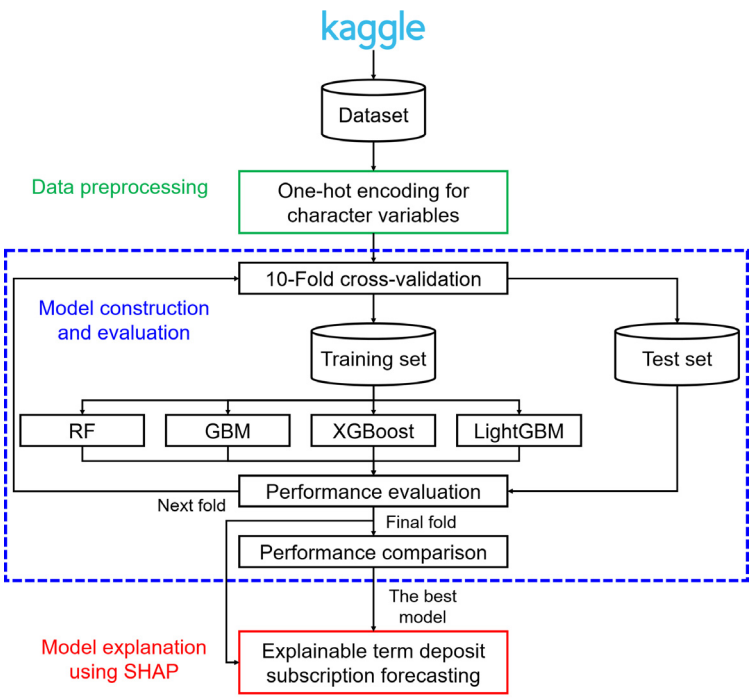
비록 이전 연구에서 개발한 ANN 기반의 분류 모델들은 만족스러운 분류 정확도를 보였지만, 블랙박스와 같이 독립변수들의 영향도와 모델의 의사결정 과정을 파악하기 어렵다[1, 12]. 인공지능 기반의 분류 또는 예측 모델이 항상 정답만 도출하는 것은 아니므로, 부정확한 예측을 수행할 때 그 이유를 파악할 수 있는 근거가 필요하다. 하지만, 대다수의 인공지능 모델은 그 근거를 확인하기가 쉽지 않아 모델의 신뢰성을 확보하는 데 한계를 보이며[29], 이로 인해 실제 산업에서 적용하기가 어려울 수 있다.

앞서 기술한 문제를 해결하기 위해, 최근 설명 가능한 인공지능(Explainable AI, XAI)에 대한 중요성이 증대되고 있으며, 이와 관련된 연구들이 진행되고 있다[1]. 금융 분야에서 신한은행[6]은 설명 가능한 신용평가를 위해, HELOC(Home Equity Line of Credit), Lending Club, UCI의 Default of Credit Card Clients 데이터 셋을 활용하여 GBM(Gradient Boosting Machine), XGBoost(eXtreme Gradient Boosting)를 이용하여 모델을 구성하였으며, SHAP (SHapley Additive exPlanations)을 각 모델에 적용하여 개인의 신용 등급을 평가하는데 어떤 독립변수가 중요한지를 분석하였다. 그 외에도 <Table 1>과 같이 다양한 분야에서 설명 가능한 인공지능을 위해 국내 연구기관에서도 SHAP을 이용한 많은 연구가 수행되었다.

비록 여러 분야에서 XAI 관련 연구들이 진행되었지만, 금융 분야 및 정기예금 가입 여부 예측에 관해 심층적인 XAI 기법의 적용 사례는 매우 미미하다. 특히 금융 분야는 클라우드 서비스 도입과 함께 AI 윤리 준칙, 데이터 내의

〈Table 1〉 Summary of Previous Works Based on XAI with SHAP

Ref.	Category	Area	Purpose	AI Method
[6]	Classification	Economic	Personal credit rating classification to provide reasons for its change	GBM, XGBoost
[7]	Multi-class classification	Economic	Classification of sentence data that explain the rationale for the rise and fall of the stock prices	Attention-based bidirectional-long short-term memory (Bi-LSTM)
[8]	Classification	Business	Process mining to improve a service refund process	DT, RF, multinomial naïve bayes, logistic regression
[11]	Prediction	Education	Students′ final grade prediction in mathematics	XGBoost
[15]	Prediction	Energy	Hourly electrical load forecasting	XGBoost
[18]	Classification	Architecture	Failure mode and effects analysis of reinforced concrete (RC) members	RF
[24]	Classification	Safety engineering	Occupational accident prediction	LightGBM
[28]	Prediction	Pathology	Influenza occurrence prediction	LightGBM
[29]	Classification	Energy	Anomaly detection for differential pressure control valve	RF



〈Figure 1〉 System Architecture

개인정보보호, 금융회사 보안 등으로 설명력 확보에 대한 외부 규제가 점차 강화되고 있다 [6, 16, 31, 37]. 따라서 금융사 임직원과 고객들에게 AI 모델의 예측값에 관하여 충분한 설명을 제공하여 신뢰성을 확보하기 위해 XAI의 필요성이 증대되고 있다.

본 논문은 AI 모델의 신뢰성을 확보하기 위해, 정기예금 가입 여부 예측에서 고객 정보의 영향도와 의사결정 과정을 해석할 수 있는 설명 가능한 분류 모델을 제안한다. 이를 위해, 먼저 테이블 형식의 데이터(Tabular Data)에서 우수한 예측 성능을 도출하는 DT 기반의 앙상블 학습 기법들을 이용하여 4가지 분류 모델들을 구성한다. 다음으로 가장 우수한 성능을 보인 모델을 대상으로 XAI 기법인 SHAP[17]을 적용하여 설명 가능한 정기예금 가입 여부 예측을 수행한다.

본 논문의 주요 기여도는 다음과 같다.

1. 본 논문은 DT 기반 앙상블 학습 기법인 랜덤 포레스트, GBM, XGBoost, LightGBM (Light Gradient Boosting Machine)을 기반으로 분류 모델들을 구성하고, 모델의 분류 성능을 Accuracy, Kappa, F1-Score 측면에서 심층 분석한다.
2. 본 논문은 가장 우수한 분류 성능을 도출하는 모델과 Lee and Hwang[14], Park et al.[27]에서 개발한 ANN 모델과의 분류 성능을 비교하여 본 논문에서 제안한 기법의 효용성을 입증한다.
3. 본 논문은 가장 우수한 예측 성능을 도출하는 모델에 XAI 기법인 SHAP을 적용하

여 고객 정보의 영향도와 의사결정 과정 등을 해석할 수 있는 근거를 제공한다.

본 논문의 나머지 구성은 다음과 같다. 제2장에서는 설명 가능한 정기예금 가입 여부 예측 모델 구성을 위한 전반적인 과정을 기술한다. 제3장에서는 제안한 기법의 실용성과 타당성을 입증하기 위한 실험 및 결과에 대해 자세히 기술한다. 마지막으로 제4장에서는 향후 연구와 함께 본 논문의 결론은 맺는다.

2. 설명 가능한 정기예금 가입 여부 예측 기법

본 논문에서 제안하는 설명 가능한 정기예금 가입 여부 예측 모델을 위한 구성도는 <Figure 1>과 같다.

2.1 데이터 전처리

본 논문은 정기예금 가입 여부 예측을 위한

〈Table 2〉 List of Independent and Dependent Variables and Their Data Type

Independent Variables	Data Type
Age	Integer
Job: type of job	Character: admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services
Marital: marital status	Character: married, divorced (divorced or widowed), single
Education	Character: unknown, secondary, primary, tertiary
Default: has credit in default?	Logical: yes or no
Balance: average yearly balance (unit: €)	Integer
Housing: has a housing loan?	Logical: yes or no
Loan: has a personal loan?	Logical: yes or no
Contact: contact communication type	Character: unknown, telephone, cellular
Day: last contact day of the month	Integer
Month: last contact month of the year	Character: Jan., Feb., Mar., …, Nov., Dec.
Duration: last contact duration (unit: sec)	Integer
Campaign: number of contacts performed during this campaign and for this client	Integer (includes last contact)
Pdays: number of days that passed by after the client was last contacted from a previous campaign	Integer (“-1” means the client was not previously contacted)
Previous: number of contacts performed before this campaign and for this client	Integer
Poutcome: outcome of the previous marketing campaign	Character: unknown, other, failure, success
Dependent Variable	Data Type
Deposit: has the client subscribed to a term deposit?	Logical: yes or no

분류 모델을 구성하기 위해, Kaggle에서 Bank Marketing Dataset[22]을 수집하였다. Bank Marketing Dataset[22]은 포르투갈의 은행에서 <Table 2>와 같이, 고객의 개인 인적 사항과 텔레마케팅 정보를 독립변수, 정기예금 가입 여부를 종속변수로 구성된 11,163개의 튜플로 이루어져 있다.

수집한 데이터 셋의 범주형 데이터는 모델 학습을 위해, 문자형(Character) 변수들은 원-핫 인코딩(One-Hot Encoding)을 수행하여 해당 속성이 속하면 1, 그렇지 않으면 0으로 관련 변수들을 생성하였으며, 논리(Logical) 또는 부울(Boolean) 변수들은 Yes는 1, No는 0으로 변숫값을 변경하였다. 이를 통해, 총 48개의 독립변수를 이용하여 종속변수인 Deposit(가입: 1, 미가입: 0)을 예측하기 위한 앙상블 학습 기반의 분류 모델들을 구성하였다.

2.2 정기예금 가입 여부 예측 모델

본 논문은 테이블 형식에서 우수한 성능을 도출하는 앙상블 학습 기법[2, 17]을 이용하여 정기예금 가입 여부 예측 모델을 구성하였다. 앙상블 학습은 여러 단일 모델을 전략적으로 생성하고 결합하여, 단일 모델보다 우수한 분류 및 예측 성능을 도출하는 방법론이다[33, 35]. 대표적인 앙상블 학습 방법론으로 배깅(Bagging), 부스팅(Boosting), 스태킹(Stacking)이 있으며[35], 스태킹은 여러 단일 모델을 사용하는 배깅, 부스팅과는 다른 학습 방식으로 여러 이기종 모델을 구성한 뒤, 각 모델의 예측값을 다시 분류 또는 예측 모델이 학습하여 최종값을 도출하는 방식이다[19].

본 논문은 분류 모델의 해석을 용이하게 하기 위해 여러 이기종 모델들을 구성해야 하는

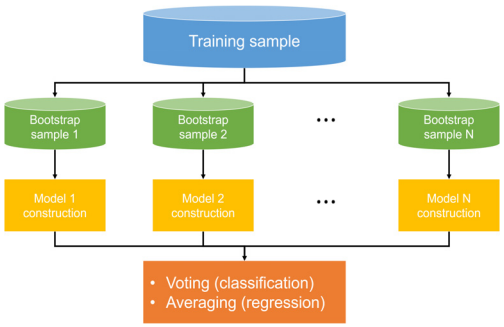
〈Table 3〉 Advantage and Disadvantage of Tree-based Ensemble Learning Methods

Methods	Advantage	Disadvantage
RF	- Can effectively handle a large number of independent variables without variable deletion - Good predictive performance with relatively little hyperparameter tuning	- Building and testing the model is somewhat slower than other machine learning methods - Unsatisfactory performance on high dimensional sparse data
GBM	- Good predictive performance for ranking or Poisson regression which RF is harder to achieve - Several hyperparameter tuning options make the function very flexible	- Building and testing the model is considerably slower than other machine learning methods - Can overemphasize outliers and cause overfitting
XGBoost	- Can reduce overfitting based on regularization technique and train more quickly than GBM or RF - Good predictive performance on small data, data with subgroups, big data, and complicated data	- Unsatisfactory performance on sparse data and very dispersed data - Prone to overfitting if hyperparameters are not adjusted correctly
LightGBM	- Faster training speed, higher efficiency, better accuracy than another popular boosting algorithm - Support of parallel or GPU learning and lower memory usage	- Prone to overfitting when trained on small datasets (< 10,000 records)

스태킹 방법론을 고려하지 않고, SHAP 패키지에서 제공하는 배깅과 부스팅 방법론만 고려하였다. <Table 3>과 같이 총 4가지의 앙상블 학습 기법을 선정하였으며, 배깅 방법론은 랜덤 포레스트, 부스팅 방법론은 GBM, XGBoost, LightGBM으로 선정하였다.

2.2.1 배깅(Bagging)

배깅[3]은 부트스트랩(Bootstrap)과 Aggregating의 합성어로 부트스트랩은 표본 분포를 구하기 위해 데이터를 여러 번 복원 추출(Resampling)하는 것을 의미하며, 복원 추출된 데이터의 분포는 표본 분포와 근사하게 된다. 배깅의 과정은 <Figure 2>와 같다.



〈Figure 2〉 Flowchart of Bagging Process

먼저, 주어진 데이터 셋을 무작위로 복원 추출하여 분할된 각 데이터 셋에서 병렬적으로 DT와 같은 단일 모델을 구성한다. 다음으로, 단일 모델들의 예측값을 집계하여, 분류는 투표(Voting), 회귀는 평균(Average)으로 집계한다. 배깅은 여러 경우의 데이터 셋을 학습하여 최종값을 도출하기 때문에 낮은 분산으로 인해 이상치에 강인하다[3, 30].

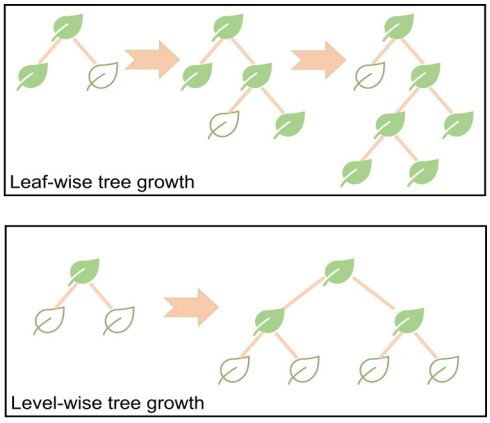
랜덤 포레스트[4, 25]는 대표적인 배깅 기법으로 독립변수를 임의로 선택하여 학습하는 방식이다. 랜덤 포레스트는 다양한 분류 및 예측 문제에서 사용되었으며[4, 20], 초매개변수(Hyperparameter)가 다른 인공지능 기법보다 적음에도 불구하고 우수한 성능을 도출한다는 장점이 있다. 대표적인 초매개변수로 나무의 수(Number of Trees)와 모델 학습을 위한 독립변수의 수(Number of Features)가 있다[21].

2.2.2 부스팅(Boosting)

부스팅은 최종 예측값을 집계하는 방식은 배깅과 유사하나, 병렬 학습인 배깅과는 다른 학습 방식으로, 단일 모델의 잔차를 다음 단일 모델이 학습하여 잔차를 줄이는 학습 방식으로[23, 30], 정교한 예측 성능을 기대할 수 있지만, 오랜 학습 시간 및 과적합(Overfitting) 문제가 발생할 수 있다[23]. GBM[23]은 부스팅 방법론의 일종으로, Basic Exact Greedy 알고리즘으로 무작위로 샘플링된 데이터에 관해 DT 모델을 구성하고, 경사 하강법(Gradient Descent)을 통해 잔차를 점차 줄여 모델을 재구성한다.

XGBoost[5]는 GBM의 오랜 학습 시간 및 과적합 문제를 해결하기 위해 개발된 인공지능 기법이다. XGBoost는 분산환경에서의 데이터 처리 문제를 효과적으로 다루기 위해 Approximate 알고리즘을 기반으로 데이터 정렬 및 분할을 통해 병렬처리를 수행하여 학습 속도를 향상하였으며, 정규화(Regularization)를 적용해 모델 구조를 간소화하여 과적합 문제를 해결하였다[5, 30]. LightGBM[10, 34]은 GBM, XGBoost보다 더욱 빠른 학습 속도와 정확도 개선을 위해 개발되었다. LightGBM은 <Figure 3>과 같이 GBM과 XGBoost에서 나무를 확장

하는 방식인 Level-Wise 방식이 아닌, Leaf-Wise 방식을 채택하여, Level-Wise에서 동일한 Level로 재조정하는 시간이 소요되지 않아 빠른 학습이 가능하다[10, 26].



〈Figure 3〉 Illustration Demonstrating the Difference between Leaf-Wise Growth (above) and Level-Wise (below)

2.2.3 SHAP 기반의 분류 모델 해석

본 논문은 4가지 앙상블 학습 모델들의 분류 성능을 심층 분석하고, 가장 우수한 분류 성능을 도출하는 분류 모델에 SHAP 기법을 적용하여 설명 가능한 정기예금 가입 여부 예측을 수행하였다.

SHAP은 섀플리 값(Shapley Value)을 기반으로 모델 학습에서 독립변수들과 종속변수의 상관관계를 효과적으로 분석할 수 있는 도구이다[17]. 섀플리 값은 게임이론에서 게임 참여자들로부터 얻어지는 모든 이득을 각 참여자의 기여분에 따라 분배된 값을 뜻한다. 예를 들어 DT 기반의 앙상블 학습 모델의 변수 중요도에서 게임은 관측치의 예측값, 참여자는 독립변수, 기여분은 변수 중요도라 가정할 수 있다.

$$\varnothing_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup i}(x_{S \cup i}) - f_S(x_S) \right] \quad (1)$$

섀플리 값은 식 (1)을 따르며, F는 모든 독립변수의 집합, f는 학습 모델, i는 독립변수를 의미한다. 또한, $\varnothing_i$는 기여분이며, $S \subset F$를 만족하는 모든 독립변수의 집합 S에 관하여 해당 독립변수인 i가 있을 때와 없을 때의 차이인 $f_{S \cup i}(x_{S \cup i}) - f_S(x_S)$를 기여분으로 산출한다. 여기서 기여분을 각 부분 집합 S에 관해, $|F|!$(모든 집합 F를 줄 세우는 경우의 수)를 $|S|!$(S를 줄 세우는 경우의 수)와 $(|F|-|S|-1)!$(i와 S를 제외한 나머지를 줄 세우는 경우의 수)으로 가중평균한다.

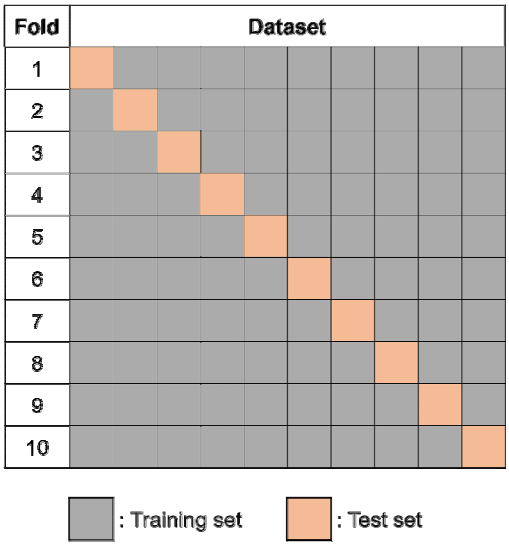
SHAP은 학습 모델의 섀플리 값을 조건부 기댓값 함수를 적용하여 값을 구하며, 모든 기계학습 모델에 적용할 수 있는 Kernel SHAP, 신경망 모델에 적용할 수 있는 Deep SHAP, 그리고 본 논문에서 적용한 DT 기반의 앙상블 학습 모델에 적용할 수 있는 Tree SHAP이 있다. Tree SHAP은 데이터 셋의 튜플마다 모든 독립변수의 SHAP 값을 산출하여 변수 중요도와 부분 의존도 그림(Partial Dependence Plot, PDP)을 제공한다[15]. 여기서 변수 중요도는 SHAP 값 또는 SHAP의 절댓값 합의 평균을 통해 그래프로 나타낼 수 있으며, PDP는 각 튜플이 갖는 해당 독립변수의 입력값과 그에 따른 SHAP 값을 모든 튜플에 관해 점으로 나타낸다. 그리하여 본 논문에서는 분류 모델의 학습 과정에 대한 예측값과 특정 독립변수가 없는 모델의 예측값과의 비교를 통해 각 독립변수의 중요도를 분석 및 시각화할 수 있다.

3. 실험 및 결과

본 논문은 Intel® Core™ i5-7600 CPU @ 3.50GHz와 16GB DDR4 RAM으로 구성된 컴퓨터 환경에서 Python 3.7.9와 PyCharm 2020.3.3의 프로그래밍 개발 도구를 통해 랜덤 포레스트, GBM 및 성능 평가 지표로는 scikit-learn 0.24.2, XGB는 xgboost 1.4.2, LightGBM은 lightgbm 3.2.1의 라이브러리를 사용하였으며, Kaggle에서 수집한 1,02MB의 크기를 가진 원본 데이터를 전처리하여 1,13MB의 크기를 가진 데이터 셋을 구성한 다음 실험을 진행하였다.

3.1 평가 방법

본 논문은 분류 모델의 성능을 심층 분석하기 위해 <Figure 4>와 같이 10겹 교차검증(10-Fold Cross-Validation)[36]을 적용하였다. 10겹 교차검증은 모델의 성능을 분석하기 위해 데이터 셋을 모두 활용하여 검증하므로 모델의 신뢰성을 높일 수 있다는 장점이 있다. 주어진 데이터 셋에서 비복원 추출 방식을 통해 총 10개의 임의의 데이터 셋을 추출한다. 추출한 임의의 데이터 셋은 1부터 10까지 모든 임의의 데이터 셋에 대해 검증하며, 검증할 임의의 데이터 셋을 제외한 나머지 9개의 임의의 데이터 셋으로 모델을 학습하는 방식으로 총 10번을 반복 검증한다.



〈Figure 4〉 Example of 10-fold Cross-Validation

3.2 평가 지표

본 논문은 분류 모델의 성능 평가 지표로 정확도(Accuracy), 카파 상관계수(Kappa Coefficient), F1-Score를 고려하였으며, 모든 평가 지표들은 <Table 4>와 같이 혼동행렬을 기반으로 평가한다.

〈Table 4〉 Confusion Matrix

Actual \ Predicted	True	False
True	TP	FN
False	FP	TN

정확도는 모든 관측치의 관측값과 모델의 예측값이 일치하는지를 나타내는 비율이다. 예를 들어, 모델의 정확도가 90%라면, 100개의 관측치 중에서 90개만 정확하게 분류한다는 의미이다. 정확도는 직관적이고 단순하다는 장점이 있으며, 식 (2)를 따른다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

카파 상관계수는 일반적으로 코헨의 카파 상

관계수를 가리키며, 관측값과 예측값과의 일치도를 측정하는 방법이며, 식 (3)부터 식 (6)까지를 따른다. 카파 상관계수의 등급은 <Table 5>와 같이 Landis와 Koch의 해석[13]을 주로 따른다.

〈Table 5〉 Guidelines of Landis and Koch

Kappa Statistic	Strength of Agreement
< 0	Poor
0 — 0.2	Slight
0.2 — 0.4	Fair
0.4 — 0.6	Moderate
0.6 — 0.8	Substantial
0.8 — 1	Almost perfect

$$P(Y) = \frac{(TP+FN)\times(TP+FP)}{(TP+TN+FP+FN)^2} \quad (3)$$

$$P(N) = \frac{(TN+FN)\times(TN+FP)}{(TP+TN+FP+FN)^2} \quad (4)$$

$$P(E) = P(Y) + P(N) \quad (5)$$

$$K = \frac{Accuracy - P(E)}{1 - P(E)} \quad (6)$$

F1-Score는 카파 상관계수와 마찬가지로 <Table 4>와 같이 혼동행렬을 기반으로 식 (7)부터 (9)까지 따르며, 정확도(Precision)와 재현율(Recall)의 조화평균을 이용하여 분류 모델의 성능을 평가하는 지표이다.

$$Precision = \frac{TP}{(TP+FP)} \quad (7)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (8)$$

$$F_1 = 2\times\frac{(\mathrm{Precision}\times Recall)}{(\mathrm{Precision}+Recall)} \quad (9)$$

3.3 모델 성능 평가

대부분의 DT 기반 앙상블 학습 기법들은 모델 성능에 많은 영향을 주는 초매개변수가 포함되

〈Table 6〉 Selected Hyperparameters for Each Tree-based Ensemble Learning Model

Methods	Ref.	Selected Hyperparameters
RF	[21]	Number of trees (*n_estimators*): 128 Number of features (*max_features*): auto, **sqrt**, log2
GBM	[26]	Number of trees (*n_estimators*): **100**, 250, 500 Learning rate (*learning_rate*): 0.01, 0.05, **0.1** Maximum depth of the individual regression estimators (max_depth): **5**, 10
XGB	[26]	Specify which booster to use (*booster*): **gbtree**, dart Number of trees (*n_estimators*): **100**, 250, 500 Maximum depth of the individual regression estimators (*max_depth*): 6, 8, **10** Subsample ratio of the training instance (*subsample*): 0.5, 0.75, **1**
LightGBM	[26]	Specify which booster to use (*booster*): gbdt, **dart** Number of trees (*n_estimators*): **1000**, 1500 Learning rate (*learning_rate*): **0.01**, 0.05, 0.1 Maximum tree leaves of the individual regression estimators (*num_leaves*): **64** Subsample ratio of the training instance (*subsample*): **0.5** Subsample ratio of columns when constructing each tree (*colsample_bytree*): **1**

어 있다. 본 논문은 <Table 6>을 참고하여 평가 방법에서 설명한 10겹 교차검증을 통해 최적의 앙상블 학습 기반 분류 모델을 구성하였다. 모든 모델을 구성할 때 Random State는 42로 설정하였다. <Table 6>에서 굵게 표시된 값은 scikit-learn에서 제공하는 GridSearchCV와 KFold 모듈에서 가장 우수한 성능을 도출하는 모델의 초매개변수 값을 나타내며, 해당 분류 모델을 기준으로 <Table 7>에 성능을 나타내었다.

〈Table 7〉 Performance Comparison (the values in the bold font indicate the best values for respective metrics)

Methods	Accuracy	Kappa	F1-Score
RF	0.857	0.715	0.850
GBM	0.863	0.726	0.860
XGBoost	0.855	0.710	0.851
LightGBM	**0.864**	**0.727**	**0.861**
ANN_1 [14]	0.539	0.078	0.524
ANN_2 [14]	0.539	0.078	0.524
DNN [27]	0.511	0.032	0.626

또한, 본 논문은 앞서 기술한 최근 문헌[14, 27]과의 모델 성능을 비교하기 위해, 최근 문헌에서 제시한 ANN 모델 구조와 동일한 분류 모델을 개발하였다. ANN의 모델은 scikit-learn에서 제공하는 MLPClassifier 모듈을 사용하였다. DT 기반의 앙상블 학습 모델과의 동일한 실험 환경을 구축하기 위해, Random State는 42로 설정하고 10겹 교차검증을 통해 모델 성능을 평가하였다. Lee and Hwang[14], Park et al.[27]에서는 데이터 전처리 과정을 자세히 기술하지 않았으므로, 본 논문에서 사용한 데이터로 표준화(Standardization)를 한 후, 모델 학습을 수행하였다. 또한, 해당 논문에서 기재하지 않은 나머지 초매개변수들은 scikit-learn의 기본값(Default Value)으로 설정하였다.

ANN_1[14]은 최적화 알고리즘으로 SGD (Stochastic Gradient Descent), 180개의 노드의 수를 갖는 1층의 은닉층, 학습률 변화를 Constant 방식으로 학습된 모델이다. ANN_2[14]는 학습률 변화만 Adaptive 방식이며, 나머지 모델 구조는 ANN_1과 같다. DNN[27]은 각 64, 128, 64개의 노드의 수를 갖는 3층의 은닉층으로 구성된 심층 신경망 모델이다.

Bank Marketing Dataset에서 10겹 교차검증을 통한 실험 결과, LightGBM이 ANN 모델들을 포함하여 모든 평가 지표에서 가장 우수한 성능을 도출하였다.

비록 신경망 모델들은 다양한 분야에서 우수한 성능을 도출하였으나, 해당 데이터 셋에서는 다소 불만족스러운 성능을 도출하였다. 이는 여러 초매개변수를 기본값으로 설정하여 실험하였기 때문에 해당 데이터 셋에 관해 적절한 학습이 어려웠음을 판단할 수 있다. 따라서 최적의 신경망 모델을 구성하기 위해서는 많은 시간이 필요하다는 것을 알 수 있다. 또한, 신경망 모델은 DT 기반의 앙상블 학습 모델보다 설명하기가 어렵다는 단점이 있으므로, DT 기반의 앙상블 학습 모델이 Bank Marketing Dataset과 같은 테이블 형식의 데이터에서 더욱 적절하다는 것을 확인할 수 있었다.

Duration(통화시간)은 실제 금융기관에서 향후 정기예금 가입을 할 잠재 고객을 찾는 목적으로 확인하기 어려운 데이터(Unseen Data)에 속하므로 실제 모델을 구성할 때 적용하기가 어렵다. 따라서, 학습 데이터 셋에서 Duration을 제외한 독립변수들을 이용하여 분류 모델을 구성하였으며, <Table 8>에 분류 모델의 성능을 나타내었다.

〈Table 8〉 Performance Comparison on the Dataset Excluding Duration (the values in the bold font indicate the best values for respective metrics)

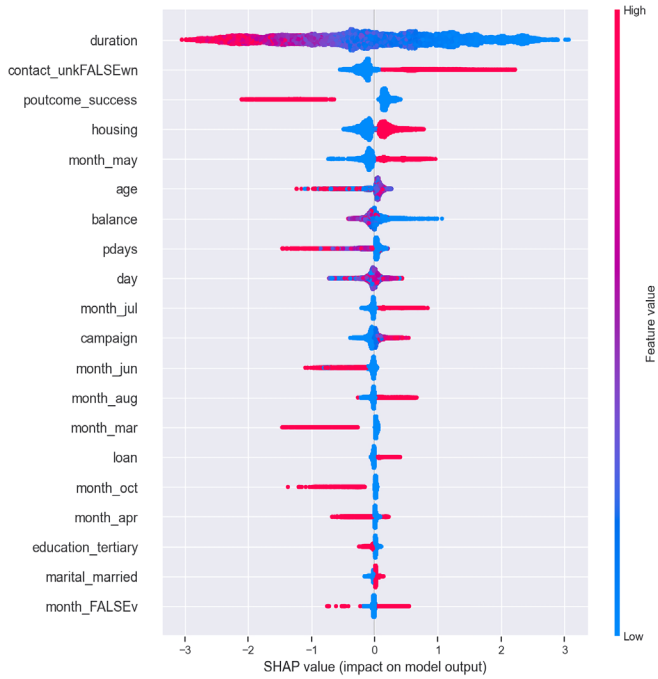
Methods	Accuracy	Kappa	F1-Score
RF	0.729	0.452	**0.694**
GBM	**0.738**	**0.469**	0.693
XGBoost	0.729	0.453	0.690
LightGBM	0.737	0.466	0.690
ANN_1 [14]	0.474	-0.062	0.265
ANN_2 [14]	0.474	-0.062	0.265
DNN [27]	0.470	-0.071	0.209

실험 결과, Duration을 제외한 데이터 셋에서 학습한 모델들은 Duration을 포함한 데이터 셋보다 전반적으로 분류 성능이 저하되었다는 것을 확인할 수 있었다. 또한, LightGBM이 가장 우수한 성능을 도출하였으나, Duration을 제외한 데이터 셋에서는 GBM이 가장 우수한 성능을 도출하였다.

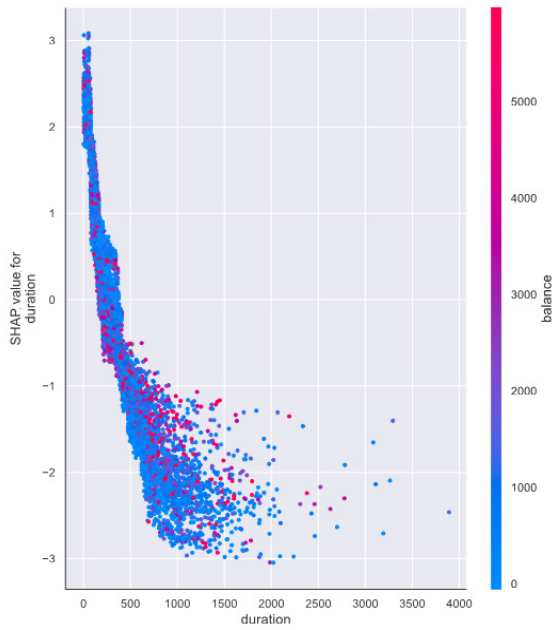
3.4 분류 모델 해석

본 논문은 먼저 Duration을 포함한 데이터 셋을 기준으로 10번째 교차검증을 수행할 때 구성된 LightGBM 모델에 SHAP을 적용하였으며, <Figure 5>부터 <Figure 8>까지 모델 해석을 위한 시각화를 수행하였다. 모든 Figure에서 제시된 독립변수에 관한 설명은 <Table 2>를 통해 확인할 수 있다.

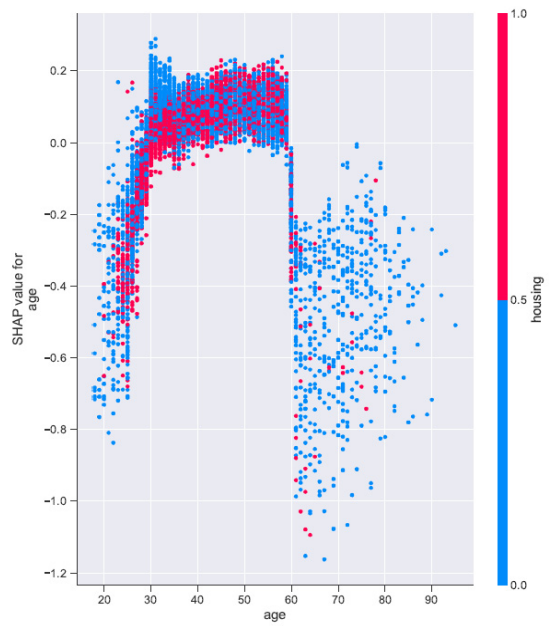
<Figure 5>는 SHAP 값들을 통해 분류 모델 구성에 영향이 큰 독립변수들을 상위 20개까지



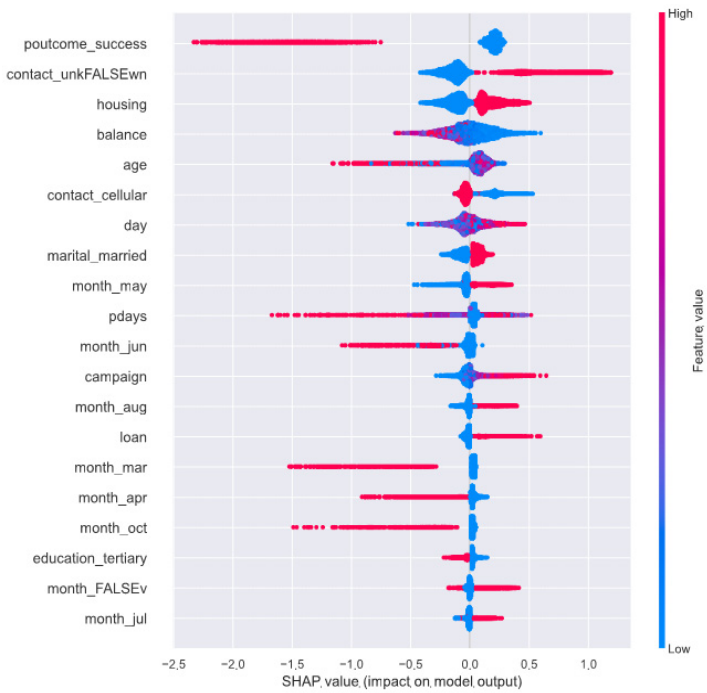
〈Figure 5〉 Summary Plot of Selected SHAP Values



〈Figure 6〉 SHAP Dependent Plot Shows that Effect of Duration on Model Output with Respect to Balance



〈Figure 7〉 SHAP Dependent Plot Shows that Effect of Age on Model Output with Respect to Housing



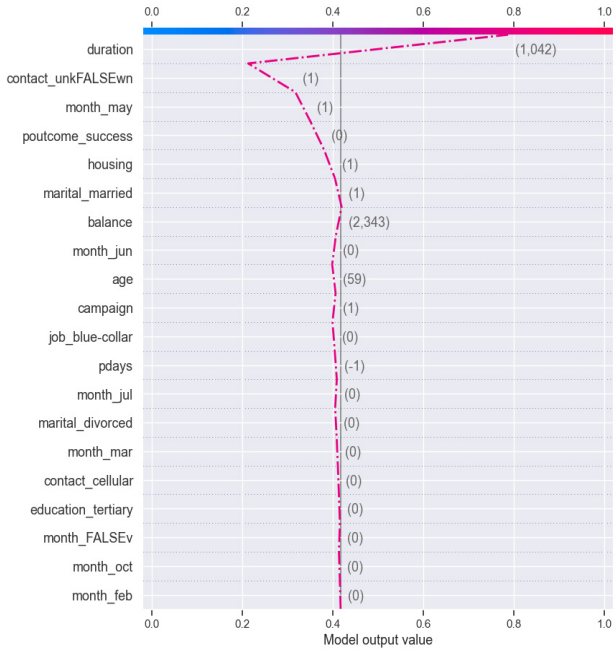
〈Figure 8〉 Summary Plot of Selected SHAP Values on the Dataset Excluding Duration

의 변수 중요도와 변수 효과를 함께 나타낸 것이다. X축은 SHAP 값의 수치, Y축은 독립변수, 색은 독립변수의 상대적인 크기를 나타낸다. 겹치는 점이 Y축 방향으로 나타남에 따라 독립변수에 관한 SHAP 값의 분포를 확인할 수 있다. SHAP 값이 양수와 음수는 정기예금 가입에 긍정적인 기여와 부정적인 기여를 각각 의미한다.

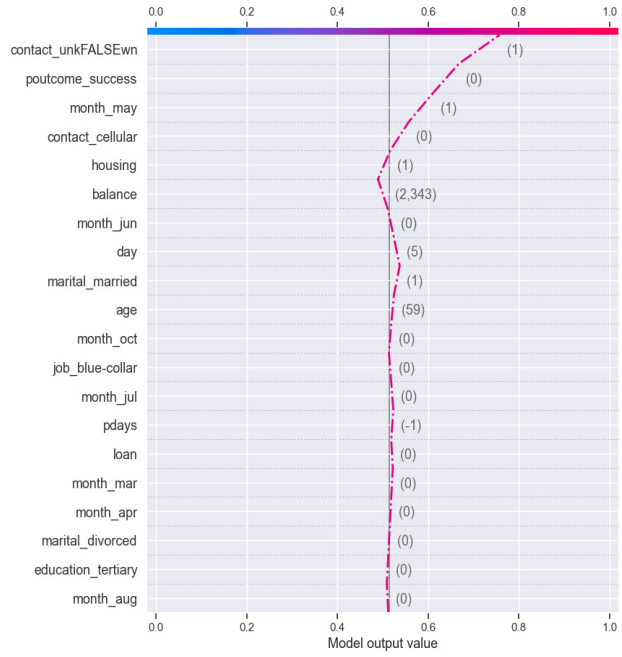
가장 큰 기여를 준 Duration(통화시간)을 살펴보면, 통화시간이 길어질수록 더욱 붉은색을 나타내며, 이는 SHAP 값이 음의 방향으로 분포되었다는 것을 알 수 있다. 이를 통해 마케팅을 위한 통화시간이 길어질수록 정기예금 가입에 부정적으로 기여한다고 해석할 수 있다.

Contact_unFalsewn은 Contact(연락수단)의 결측치를 원-핫 인코딩으로 나타낸 독립변수이다. 붉은색은 마케팅에서 고객의 연락수단을 확인하기가 어려운 경우이며, 푸른색은 일반전화 또는 휴대전화로 연락한 경우이다. SHAP 값이 양의 방향으로 분포하고 있으므로, 일반전화 또는 휴대전화가 아닌 다른 연락수단이 오히려 예금 가입에 긍정적으로 기여한다고 해석할 수 있다.

Poutcome은 이전 마케팅에 참여하여 상품에 가입하였는지를 판단하는 독립변수이다. 붉은색은 이전 마케팅을 통해 상품에 가입한 경우이며, 푸른색은 상품에 가입하지 않은 경우를 의미한다. 해당 변수의 SHAP 값을 비교하였을 때, 이전 마케팅을 통해 상품에 가입하였을 때에는 부정적인 경향을 보였으나, 상품에 가입하지 않았을 때는 SHAP의 값이 0.5 미만으로 조금이라도 긍정적인 경향을 끼친다는 것으로 해석할 수 있다.



〈Figure 9〉 SHAP Decision Plot



〈Figure 10〉 SHAP Decision Plot on the Dataset Excluding Duration

Housing은 고객이 주택담보대출을 받았는지에 관한 여부를 나타낸 것이며, 붉은색으로 표기한 주택담보대출을 받은 고객은 정기예금 가입에 긍정적인 반응을 보인다고 해석할 수 있다. Age는 고객의 나이를 나타내는 독립변수로, 붉은색과 푸른색은 각각 고령층과 청년층을 의미한다. SHAP 값의 분포를 통해 고령층과 청년층은 정기예금 가입에 부정적인 경향을 보인 반면에, 중년층은 정기예금 가입에 긍정적인 경향을 보인다는 것을 확인할 수 있었다.

Balance는 고객의 자산을 나타내는 독립변수로 자산이 많은 고객보다 자산이 적은 고객이 정기예금 가입에 긍정적으로 기여한다고 해석할 수 있었다. Pdays는 고객이 이전 마케팅에서 상품에 가입한 시점으로부터 현재까지의 시간을 나타내는 독립변수이다. 여기서, 이전 마케팅에서 상품에 가입하지 않은 고객은 -1로 푸른색을 의미한다. 이는 신규 고객으로 정기예금 가입에 긍정적인 경향을 보인다는 것을 확인할 수 있었으며, 반면에 이전 마케팅에서 금융상품에 가입한 기존 고객은 정기예금 가입에 다소 부정적인 경향을 보인다는 것을 확인할 수 있었다.

PDP는 SHAP의 Summary Plot에서 파악하기 어려운 입력변수의 값이 예측에 미치는 영향을 효과적으로 파악할 수 있다. <Figure 6>은 Balance를 기준으로 Duration(X축)에 관한 SHAP 값(Y축)의 분포를 PDP로 나타낸 것이며, SHAP 값이 0 이상일 때에는 정기예금 가입에 긍정적이다고 해석할 수 있다. Duration의 값이 500 이상부터는 SHAP의 값이 모두 음수로 나타났으며, 이는 상담시간이 길수록 정기예금에 가입하지 않았다고 해석할 수 있다.

<Figure 7>은 Housing을 기준으로 Age(X축)에 관한 SHAP 값(Y축)의 분포를 PDP로 나타낸 것이다. 20대까지는 주택담보대출 여부와 관계없이, SHAP의 값이 0 이하를 보였으며, 30세부터 60세까지의 구간에서 SHAP의 값이 0.2로 긍정적인 경향을 확인할 수 있었다. 또한, 60세를 경계로 SHAP의 값이 음수 구간을 보였으며, 이는 정기예금 가입에 부정적인 반응을 보인다는 것을 확인할 수 있었다.

앞서 <Figure 5>에서 확인할 수 있듯이, Duration은 모델 구성에 매우 큰 영향을 주는 독립변수이지만, 실제 모델을 구성할 때 적용하기 어렵다. 따라서 <Table 8>과 같이 Duration을 제외한 데이터 셋에서 가장 우수한 성능을 도출한 GBM 모델에 SHAP을 적용하여 <Figure 8>을 통해 GBM 모델에게 영향을 준 상위 20개의 변수 중요도와 변수 효과를 함께 나타내었다. Duration을 제외함으로써, SHAP 값의 범위가 <Figure 8>을 기준으로 -3부터 +3에서 -2.5부터 1로 변경되었다.

<Figure 9>는 Duration이 속한 데이터 셋에서 특정 표본(고객)의 정기예금 가입 여부를 판단하는 과정을 시각화한 것이다. 분류 모델 예측값의 기준은 약 0.4로 Duration을 제외한 모든 독립변수보다 Duration이 정기예금 가입 여부를 결정하는 데 많은 영향을 미친다고 판단하였으며, 모든 변수를 고루 고려하는 것이 아닌, 특정 독립변수만으로 분류 모델의 예측값이 결정될 수 있다고 확인할 수 있었다.

<Figure 10>은 Duration을 제외한 학습 데이터 셋에서 <Figure 9>와 동일한 표본의 정기예금 가입 여부를 판단하는 과정을 시각화한 것이다. <Figure 9>와 달리 <Figure 10>은 분류 모델 예측값의 기준이 약 0.5로 여러 독립변수에 높은 가중치를 할당하여 모델의 예측값을 결정한다는 것을 확인할 수 있었다.

4. 결 론

본 논문은 설명 가능한 정기예금 가입 여부 예측을 위해 테이블 형식에서 우수한 분류 성능을 도출하는 앙상블 학습 모델인 랜덤 포레스트, GBM, XGBoost, LightGBM을 기반으로 분류 모델을 구성하였다. 실험을 위해 Kaggle에서 수집한 Bank Marketing Dataset을 이용해 10겹 교차검증을 수행하였으며, 평가 지표로 정확도, 카파 상관계수, F1-Score를 적용하였다. 실험 결과, LightGBM이 0.864의 정확도, 0.727의 카파 상관계수, 0.861의 F1-Score로 가장 우수한 성능을 도출하였으며, XAI 기법인 SHAP을 LightGBM 모델에 적용한 결과, Duration(통화 시간)이 모델 구성에서 가장 큰 영향을 끼친다는 것을 확인하였다. Duration을 제외하였을 때에는 GBM이 0.738의 정확도, 0.469의 카파 상관계수, 0.693의 F1-Score로 가장 우수한 성능을 도출하였음을 확인할 수 있었다.

또한, 본 논문은 분류 모델에 SHAP을 적용하여 분류 모델 예측값의 판단 근거를 제공하였다. 모든 독립변수 중 상위 7개인 Duration, Age(나이), Housing(주택담보대출)과 같은 특정변수들의 영향도를 분석하였다. 통화시간이 500초보다 짧을 때, 주택담보대출 또는 개인 대출을 보유하고 있을 때, 30대부터 50대까지의 고객들은 정기예금 가입에 긍정적인 반응을 보인다는 것을 확인할 수 있었다. 이뿐만 아니라 특정 고객의 정보를 SHAP 기법에 적용하여 분류 모델의 의사결정을 확인할 수 있었다. 전반적으로 Duration을 제외한 데이터 셋이 여러 독립변수를 효과적으로 고려하여 예측값을 도출한다는 것을 확인할 수 있었다.

본 연구는 국내 실제 금융 데이터를 확보하기 어려워, 포르투갈 금융기관의 데이터로 실험을 진행한 아쉬움이 있다. 향후 국내의 금융 데이터를 확보하게 된다면, 모델 성능 고도화 및 해석을 통해 국내 금융 산업에서 실질적인 도움이 되는 연구를 수행하고자 한다.

References

[1] Adadi, A. and Berrada, M., “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, Vol. 6, pp. 52138-52160, 2018.

[2] Ahmadi, A., Nabipour, M., Mohammadi-Ivatloo, B., Amani, A. M., Rho, S., and Piran, M. J., “Long-Term Wind Power Forecasting Using Tree-Based Learning Algorithms,” IEEE Access, Vol. 8, pp. 151511-151522, 2020.

[3] Altman, N. and Krzywinski, M., “Ensemble methods: bagging and random forests,” Nature Methods, Vol. 14, No. 10, pp. 933-935, 2017.

[4] Belgiu, M. and Drăguţ, L., “Random forest in remote sensing: A review of applications and future directions,” ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 114, pp. 24-31, 2016.

[5] Chen, T. and Guestrin, C., “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, pp. 785-794, 2016.

[6] Chun, Y. E., Kim, S. B., Lee, J. Y., and Woo, J. H., "Study on credit rating model using explainable AI," Journal of the Korean Data and Information Science Society, Vol. 32, No. 2, pp. 283-295, 2021.

[7] Chun, Y. E., Park, Y., Sung, N., and Choi, J., "Model analysis using estimation of shapley value on classification of sentences explaining causes of changes in stock prices," KIISE Transactions on Computing Practices, Vol. 26, No. 4, pp. 195-201, 2020.

[8] Jung, C. and Lee, H., "A comparative study of explainable AI techniques for process analysis," Journal of the Institute of Electronics and Information Engineers, Vol. 57, No. 8, pp. 51-59, 2020.

[9] Hung, P. D., Hanh, T. D., and Tung, T. D., "Term deposit subscription prediction using spark MLlib and ML packages," in Proceedings of the 2019 5th International Conference on E-Business and Applications, pp. 88-93, 2019.

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., "LightGBM: A highly efficient gradient boosting decision tree," in Proceedings of the 31st International Conference on Neural Information Processing Systems, Vol. 30, pp. 3146-3154, 2017.

[11] Kim, S., Kim, W., Jang, Y., and Kim, H., "Development of Explainable AI-Based Learning Support System," The Journal of Korean Association of Computer Education, Vol. 24, No. 1, pp. 107-115, 2021.

[12] Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., Sun, J., and Choo, J., "RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records," IEEE Transactions on Visualization and Computer Graphics, Vol. 25, No. 1, pp. 299-309, 2018.

[13] Landis, J. R. and Koch, G. G., "An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers," Biometrics, pp. 363-374, 1977.

[14] Lee, D. Y. and Hwang, B. S., "Performance comparison of algorithm for the prediction of time deposit," in Proceedings of the Korea Computer Congress, pp. 2074-2076, 2018.

[15] Lee, Y.-G., Oh, J.-Y., and Kim, G., "Interpretation of load forecasting using explainable artificial intelligence techniques," The Transactions of the Korean Institute of Electrical Engineers, Vol. 69, No. 3, pp. 480-485, 2020.

[16] Lim, M. and Jang, H., "A Study on the Risk Reduction Plan of Cryptocurrency Exchange," Journal of Platform Technology, Vol. 8, No. 4, pp. 29-37, 2020.

[17] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee,

S.-I., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, Vol. 2, No. 1, pp. 56-67, 2020.

[18] Mangalathu, S., Hwang, S. H., and Jeon, J. S., “Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach,” Engineering Structures, Vol. 219, p. 110927, 2020.

[19] Moon, J., Jung, S., Rew, J., Rho, S., and Hwang, E., “Combination of short-term load forecasting models based on a stacking ensemble approach,” Energy and Buildings, Vol. 216, p. 109921, 2020.

[20] Moon, J., Kim, J., Kang, P., and Hwang, E., “Solving the Cold-Start Problem in Short-Term Load Forecasting Using Tree-Based Methods,” Energies, Vol. 13, No. 4, p. 886, 2020.

[21] Moon, J., Kim, Y., Son, M., and Hwang, E., “Hybrid short-term load forecasting scheme using random forest and multilayer perceptron,” Energies, Vol. 11, No. 12, p. 3283, 2018.

[22] Moro, S., Cortez, P., and Rita, P., “A data-driven approach to predict the success of bank telemarketing,” Decision Support Systems, Vol. 62, pp. 22-31, 2014.

[23] Natekin, A. and Knoll, A., “Gradient boosting machines, a tutorial,” Frontiers in Neurorobotics, Vol. 7, p. 21, 2013.

[24] Oh, H. R., Son, A. L., and Lee, Z., “Occupational accident prediction modeling and analysis using SHAP,” Journal of Digital Contents Society, Vol. 22, No. 7, pp. 1115-1123, 2021.

[25] Oshiro, T. M., Perez, P. S., and Baranauskas, J. A., “How Many Trees in a Random Forest?,” Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 154-168, 2012.

[26] Park, J., Moon, J., Jung, S., and Hwang, E., “Multistep-ahead solar radiation forecasting scheme based on the light gradient boosting machine: A case study of Jeju Island,” Remote Sensing, Vol. 12, No. 14, p. 2271, 2020.

[27] Park, S. H., Lee, J. H., Jung, Y. W., and Won, Y. J., “Performance comparison of periodic deposit prediction using machine learning,” Proceedings of the Korea Software Congress, pp. 2139-2141, 2018.

[28] Park, S., Moon, J., Jung, S., Jung, S., and Hwang, E., “SHAP-based Explainable Influenza Occurrence Forecasting using LightGBM,” Proceedings of the Korea Software Congress, pp. 666-668, 2020.

[29] Park, S., Moon, J., and Hwang, E., “Explainable anomaly detection for district heating based on shapley additive explanations,” Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW), pp. 762-765, 2020.

[30] Park, S., Moon, J., Jung, S., Rho, S., and Baik, S. W., Hwang, E., “A two-stage industrial load forecasting scheme for

day-ahead combined cooling, heating and power scheduling," Energies, Vol. 13, No. 2, p. 443, 2020.

[31] Park, W. and Jang, H., "A study on implementing a priority tasks for invigoration of cloud in financial sector," Journal of Platform Technology, Vol. 8, No. 1, pp. 10-15, 2020.

[32] Parlar, T., "Using Data Mining Techniques for detecting the important features of the bank direct marketing data," International Journal of Economics and Financial Issues, Vol. 7, No. 2, p. 692, 2017.

[33] Rew, J., Cho, Y., Moon, J., and Hwang, E., "Habitat suitability estimation using a two-stage ensemble approach," Remote Sensing, Vol. 12, No. 9, p. 1475, 2020.

[34] Rew, J., Kim, H., and Hwang, E., "Hybrid segmentation scheme for skin feature extraction using dermoscopy images," Computers, Materials & Continua, Vol. 69, No. 1, pp. 801-817, 2021.

[35] Ribeiro, M. H. D. M., and dos Santos Coelho, L., "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," Applied Soft Computing, Vol. 86, p. 105837, 2020.

[36] Rodriguez, J. D., Perez, A., and Lozano, J. A., "Sensitivity analysis of k-fold cross validation in prediction error estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 3, pp. 569-575, 2009.

[37] Sun, J. C. and Kim, I. S., "Improvement of selective consent method in the collection process of personal information of financial institutions," The Journal of Society for e-Business Studies, Vol. 25, No. 1, pp. 123-134, 2020.

저 자 소 개

신지안 (E-mail: lanta4825@cau.ac.kr)
2015년~2021년 상명대학교 정보보안공학과 (학사)
2021년~현재 중앙대학교 융합보안학과 (석사과정)
관심분야 산업보안, 정보보안, 기계학습 등

문지훈 (E-mail: johnny89@cau.ac.kr)
2009년~2015년 한성대학교 정보통신공학과 (학사)
2015년~2021년 고려대학교 전기전자공학과 (박사)
2021년~현재 Topic Editor of Sustainability
2021년~현재 중앙대학교 박사후연구원
관심분야 에너지 예측, 시계열 분석, 기계학습 응용 등

노승민 (E-mail: smrho@cau.ac.kr)
2008년 아주대학교 정보통신공학과 (박사)
2008년~2009년 Postdoctoral Researcher, Carnegie Mellon University
2009년~2012년 고려대학교 전기전자전파공학부 연구교수
2012년~2013년 백석대학교 정보통신공학과 조교수
2013년~2019년 성결대학교 미디어소프트웨어학과 조교수
2019년~2021년 세종대학교 소프트웨어학과 조교수
2021년~현재 중앙대학교 산업보안학과 부교수
관심분야 빅데이터 보안, 인공지능 보안, 콘텐츠 보안 등