인공지능 알고리즘을 이용한 은행 예금 가입 의사 분류 예측에 관한 연구

황용우 [1], 박성환 [2], 김재현 [1], 신현우 [3], 조석헌*
국립금오공과대학교 [1], 한남대학교 [2], 전북대학교 [3], *University of California, San Diego (UCSD)

lmsshh88@gmail.com, 0324sunghwan@gmail.com, tkdl1014@gmail.com, 18ijustdoit@gmail.com, *justinshcho@gmail.com

Study on Prediction of Bank Deposit Subscription Intention Classification Using Artificial Intelligence Algorithms

Yongwoo Hwang[1], Sungwhan Park[2], Jaehyeon Kim[1],
Hyeonwoo Shin[3], and Seokheon Cho*
Kumoh National Institute of Technology[1], Hannam University[2], Jeonbuk National University[3], *University of California, San Diego (UCSD).

Abstract

This study presents a model based on artificial intelligence algorithms, which predicts deposit subscription intentions using bank customer data in Portuguese. The artificial intelligence algorithms employed for deposit subscription prediction model include Logistic Regression, Random Forest, and Gradient Boost Machine. To resolve data imbalance that is a critical issue in the Portuguese bank user dataset, we utilized various oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE, and Adaptive Synthetic Sampling for Imbalanced Learning (ADASYN). The objective of our provided models for predicting bank deposit subscription intentions is to accurately identify potential subscribers. A model attains better performance, as it achieves a higher recall that can result in a higher F2 score. Our analysis showed that the Gradient Boost Machine algorithm-based deposit subscription prediction model, employed by ADASYN oversampling, reaches the best performance.

Ⅰ. 서 론

은행은 유동성 확보와 수익 증대를 위해 지속적으로 고객을 대상으로 예금 가입 권유 캠페인을 실시한다 [1]. 인공지능 기술 발전과 더불어 은행은 인공지능 알고리즘을 이용하여 고객들의 상품 구매 이력, 채무 여부 및 직업과 같은 특성을 분석하여 미래의 잠재 고객들을 분류하는 업무는 4차 산업 시대의 금융환경에서 필수적인 요소이다 [2]. 따라서 본 연구에서는 은행들은 이러한 접근방식을 통해 경쟁 우위를 확보하고 은행 수익의 지속적인 성장을 도모할 수 있도록 은행 고객들의 예금 상품 가입 의사 분류 예측 모델을 제안하고자 한다.

인공지능 알고리즘을 이용한 은행 예금 가입 의사를 예측하는 선행 연구들이 아래와 같이 존재한다. 주로 포르투갈 은행 이용자에 관한 데이터 세트를 사용하고 있다 [3]. Sipu Hou *et al.*는 예금 가입 의사가 없는 사람들을 예측하는 모델을 제시하여 은행 예금 가입 의사가 없는 고객들에게 은행이 무분별하고 반복적으로 광고하여 은행의 이미지에 손상이 가는 것을 방지하였다 [4]. Random Forest (RF) 기반 모델이 0.919의 가장 높은 정확도 (accuracy)를 보였고, Artificial Neural Network (ANN) 기반 예측 모델이 0.999의 가장 높은 민감도 (sensitivity)를 가졌다. Marcus Hjertaas et al.은 잠재 고객 손실률 (Possible Customer Loss value: PCL)을 분석하였다 [5]. 고려하는 데이터 세트의 불균형을 언더샘플링 (under-sampling) 기법을 이용하여 해소하였다. 분석 결과 2.83%의 잠재 고객 손실률을 가진 Support Vector Machine (SVM) 기반 모델이 최적의 모델로 선정하였다. Much Aziz Muslim et al.의 연구는 데이터 불균형을 해소하기 위해서 Synthetic Minority Over-sampling Technique (SMOTE) 을 이용하였으며, Light Gradient-Boosting Machine (LightGBM)이 0.9063으로 가장 높은 정확도를 보여주었다 [6]. 본 연구에서는 은행 예금 가입 의사 분류 예측 모델의 성능을 향상시키기 위하여 선행 연구들과 달리 다양한 오버샘플링 (oversampling) 기법을 사용하였다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 원본 데이터 세트에 관한 내용과 전처리 과정을 서술한다. 제 Ⅲ장에서는 본 연구에 사용한 알고리즘들과 최적의 모델 선정을 위한 평가지표 그리고 데이터 불균형을 해소하기 위한 다양한 오버샘플링 기법들을 소개한다. 이후, 제 Ⅳ장에서는 제시한 은행 예금 가입 의사 분류 예측 모델들에 대한 성능을 비교 분석한다. 마지막으로 본 연구의 결과와 향후 연구 과제를 제 Ⅴ장에서 제시한다.

Ⅱ. 원본 데이터 세트 설명 및 전처리 과정

2.1 원본 데이터 세트 설명

본 연구에서는 은행 예금 가입 의사 분류 예측 모델을 위해서 포르투갈 은행 이용자에 관한 데이터를 사용하였다 [6]. 이 원본 데이터 세트는 2008년부터 2010년까지 3년간 수집한 데이터로 총 45,211개의 샘플들로 구성되어 있다. 원본 데이터 세트에 포함된 특성들과 각 변수의 데이터 형태와 가능한 값들은 표 1에 나타내었다. 본 연구에서 고려하는 데이터 세트는 다음과 같은 16개의 독립변수들이 있다. 은행 이용자의 나이 (Age), 직업 (Job), 결혼 유무 (Marital), 학력 (Education), 채무 불이행 여부 (Default), 은행의 잔액 (Balance), 주택 담보 대출 경험의 유무 (Housing), 주택 담보 대출 이외에 타 대출 경험의 유무 (Loan), 집 전화, 휴대전화과 같은 연락 수단 (Contact), 은행 예금과 관련된 연락이 발생한 달 (Month)과 일시 (Day), 마지막 연락에서 통화가 지속된 시간 (Duration), 지난 예금 가입 광고 연락이 발생하기 이전에 발생한 연락 횟수 (Campaign), 연락이 발생한 마지막 날로부터 지난 일 수 (Pdays), 해당 예금상품 이전에 예금에 가입했던 횟수 (Previous), 이전 예금상품 가입 권유의 성공 여부 (Poutcome)들이다. 여기에서, 은행의 잔액의 단위는 유로 (Euro)이고 연락이 발생한 마지막 날로부터 지난 일 수의 값이 -1이면 이전 접촉이 없었음을 의미한다. 종속변수는 예금 가입 권유 성공 여부 (Acceptance)이다. 종속변수의 값은 88%의 거절과 12%의 승낙으로 구성되어 있어본 연구에서 고려하는 데이터 세트은 불균형 데이터라 할 수 있다.

2.2 원본 데이터 세트에 대한 전처리 과정

은행에서 고객을 대상으로 하는 예금 권유는 온라인상에서 이루어지는 경우가 많은데 마케팅이 진행되는 시점에 고객의 상태를 고려해야 한다. 예를 들면, 업무가 많고 여유시간이 상대적으로 부족한 월요일부터 목요일에는 예금 권유 전화를 오래동안 받기 어려울 수 있으며, 예금에 대한 정보를 온전히 전달하지 못해 예금 권유에 대해 거부할 가능성이 높다. 이와는 달리, 금요일과 주말에는 출근이을 하지않는 고객들이 많을 뿐만 아니라 시간적 여유가 있기 때문에 상대적으로 오랜 시간동안 예금 권유 및 다양한 정보전달을 통해 마케팅을 성공으로 이끌 가능성이 높을 수 있다. 이러한 판단하에 은행 예금과 관련된 연락이 발생한 달과 일시 변수들을 이용하여 당일에 해당하는 요일 (Date) 변수를 생성 및 추가하였다.

본 연구에서 고려하고 있는 포르투갈 은행 이용자에 관한 데이터는 심한 불균형 데이터이다. 예측 모델들의 성능을 향상시키기 위해서 불균형 데이터를 해소하는 오버샘플링 기법들을 적용하였다. 다양한 오버샘플링 기법들 중에서 Synthetic Minority Over-sampling Techni-que (SMOTE), Borderline SMOTE 그리고 Adaptive Synthetic Sampling for Imbalanced Learning (ADASYN)들을 사용하였다. 하지만, 이러한 오버샘플링 기법들은 수치형 데이터에 적절한 방법이기 때문에, 원본 데이터 세트에 은행 예금과 관련된 연락이 발생한 요일 (Date) 변수가 포함된 총 17개의 독립변수들 중에서 11개의 범주형 데이터를 인코딩하여 수치형 데이터로 변환하였다. Education, Default, Housing 그리고 Loan 등과 같은 4개의 독립변수는 객체를 0과 1처럼 나눌 수 있는 데이터이므로 라벨 인코딩을 적용하였다. 그 외의 7개의 범주형 데이터는 원-핫 인코딩을 적용하였다.

표 1. 원본 데이터 세트의 구성 특성들

Independent Variables		
Variable	**Type**	**Values**
Age	Numerical [Integer]	{18, ..., 95}
Job	Categorical [String]	{Admin, Unknown, Unemployed, ⋯, Services}
Marital	Categorical [String]	{Married, Divorced, Single}
Education	Categorical [String]	{Primary, Secondary, Tertiary, Unknown}
Default	Categorical [String]	{Yes, No}
Balance	Numerical [Integer]	{-8,019, ..., 102,127} [EUR]
Housing	Categorical [String]	{Yes, No}
Loan	Categorical [String]	{Yes, No}
Contact	Categorical [String]	{Unknown, Telephone, Cellular}
Month	Categorical [String]	{Jan, ..., Dec}
Day	Categorical [Integer]	{1, ..., 31}
Duration	Numerical [Integer]	{0, ⋯, 4918} [Sec]
Campaign	Numerical [Integer]	{1, ..., 63}
Pdays	Numerical [Integer]	{-1, ..., 871}
Previous	Numerical [Integer]	{0, ..., 275}
Poutcome	Categorical [String]	{Success, Other, Failure, Unknown}
Dependent Variable		
Variable	**Type**	**Values**
Acceptance	Categorical [String]	{Yes, No}

Ⅲ. 인공지능 알고리즘 및 오버샘플링 처리

3.1 인공지능 알고리즘 및 성능 지표

본 연구에서는 Logistic Regression (LR), Random Forest (RF) 및 Gradient Boost Machine (GBM) 등 총 세 가지 인공지능 알고리즘들을 사용하여 은행 예금 가입 의사 분류를 예측하였다. LR은 종속변수가 이항 분포를 따를 때 분류 및 예측을 수행하는 알고리즘이다. 독립변수들과 가중치의 조합을 이용하여 종속변수를 분류한다. 본 연구에서 활용한 데이터 세트들에 대한 최적의하이퍼 파라미터는 다음과 같다. Epsilon 값은 1.0E-5이고 학습률은 0.05이다. RF는 의사결정 트리의 앙상블 학습 방법으로 다수의 의사 결정 트리를 생성하고 결과값을 결합하여 보다 정확도 높은 모델을 구축한다. 본 연구에서는 Gini Index를 기준치로 사용하였고, 과적합 방지를 위해 트리의 최대 깊이를 10으로 제한하였다. 앙상블 효과를 극대화하기 위해 100개의 모델 개수를 이용하였다. GBM은 다수의 약한 학습기를 순차적으로 학습시키며, 이전 단계 학습기를 활용하여 예측 오류를 줄이는 방향으로 모델을 구축하는 앙상블 학습 방법이다. 본 연구에서는 최대 트리 깊이를 4로

설정하여 복잡도를 제어하였으며, 250개의 모델 개수와 0.07의 고정 학습률을 이용하였다.

성능 평가 지표로는 정확도와 F2 score를 이용하였다. 실제로 예금 가입 의사가 있는 고객이지만 잘못 분류 예측하여 발생하는 은행의 손해를 줄이기 위해 재현율에 더욱 높은 가중치를 적용한 F2 score를 사용한다.

3.2 불균형 데이터 처리를 위한 오버샘플링 기법

본 연구에서 사용하는 데이터 세트의 불균형성을 해결하기 위해 Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE 그리고 Adaptive Synthetic Sampling for Imbalanced Learning (ADASYN) 등 총 3 가지의 샘플링 기법을 사용하였다. SMOTE은 소수 클래스 샘플들 사이에 가상의 연결선을 생성하여 그 위에 가상의 샘플을 생성하여 다수 클래스 샘플과 수를 맞추는 오버샘플링 기법이다. Borderline SMOTE은 SMOTE의 변형 중 하나로, 특히 소수 클래스 데이터들이 다수 클래스 데이터들과의 경계선 근처에 위치하는 경우에 효과적인 오버샘플링 기법이다. 다수 클래스 경계선에 위치한 소수 클래스들의 데이터들을 중점적으로 복제하는 방식이다. ADASYN은 SMOTE의 확장형 오버샘플링 기법으로, 소수 클래스 데이터들과 다수 클래스 데이터들이 섞여 있는 경계에 집중적으로 소수 클래스의 데이터들을 가상으로 생성한다는 점이 Borderline SMOTE과 공통점이다. 그러나 ADASYN은 소수 클래스 데이터들을 생성할 때 생성할 데이터의 중요도를 판단하여 생성 여부를 결정하는 점에 있어서 차별성이 있다.

Ⅳ. 은행 예금 가입 의사 분류 예측 모델 분석

4.1 학습 및 테스트 데이터 세트 구성

그림 1은 은행 예금 가입 의사 분류 예측 모델을 위해서 사용하고 있는 데이터 세트의 분할 (partitioning)과 오버샘플링 적용 과정을 보여주고 있다.

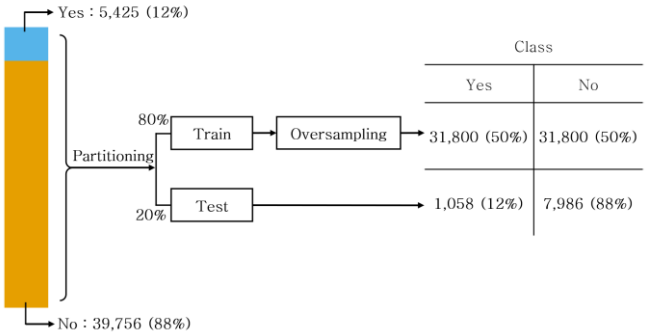

그림 1. 데이터 세트 분할과 오버샘플링 적용방법

무엇보다도 먼저, 총 45,211개의 데이터 샘플들을 학습 데이터 세트와 테스트 데이터 세트로 각각 80%와 20%로 분할하였다. 이 후 학습 데이터 세트에 대해서는 3 개의 오버샘플링 기법을 적용하여, 종속변수인 Acceptance의 2 개의 라벨인 Yes와 No의 개수가 모두 31,800개로 동일하다. 즉, 이 경우 2개의 라벨의 비율이 50%로 동일한 것이다. 이와는 달리, 테스트 데이터 세트에 대해서는 오버샘플링 기법을 적용하지 않았다. 따라서, 2개의 라벨인 Yes와 No의 개수가 각각 1,058개와 7,986개이고, 이 둘의 비율은 12%와 88%이다. 이 비율은 분할 전 종속변수의 2 개 클래스의 비율과 동일하다. 이러한 방식의 학습 및 테스트 데이터 세트 분할 후 학습 데이터 세트에만 오버샘플링 기법을 적용한 이유는 학습 시에는 소수 클래스의 데이터 개수를 늘려 학습 효율성을 증대한 반면 실제로 수집하였던 데이터를 가지고 테스트를 진행하여 예측 모델 성능의 현실성을 증가시켰다.

4.2 은행 예금 가입 의사 분류 예측 모델 성능 평가

그림 2는 본 연구에서 제시하는 은행 예금 가입 의사 분류 예측 모델들의 정확도를 보여주고 있다. 3 개의 알고리즘 기반 모델들 전반적으로 오버샘플링 기법을 적용함에 따라 오버샘플링 종류와 상관없이 정확도가 떨어지거나 동일하게 유지함을 확인할 수 있었다. 오버샘플링 적용 이후에 LR 기반 모델들의 정확도는 동일하게 유지된 반면에 RF 기반 모델들의 정확도는 감소하였다.

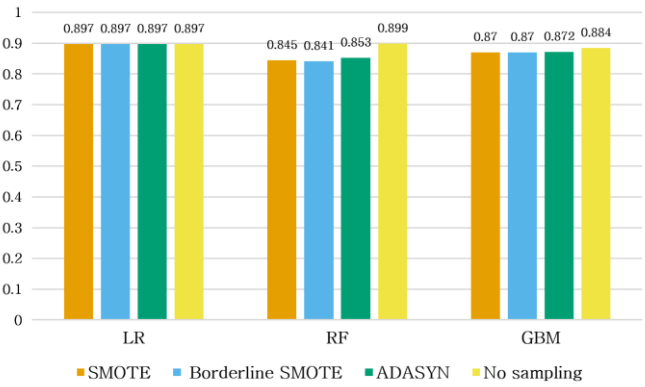

그림 2. 은행 예금 가입 의사 분류 예측 모델들의 정확도 비교

그림 3은 LR, RF 및 GBM 알고리즘 기반 모델들의 F2 score값을 나타내고 있다. 그림 2와 그림 3의 결과를 종합한다면, RF 알고리즘 기반 모델들이 오버샘플링 기법을 적용함으로써 정확도는 다소 떨어진 반면에 재현율의 가중치가 높은 F2 score값은 크게 상승함을 볼 수 있다. 또한, GBM 알고리즘을 사용하는 모델들은 오버샘플링 기법 적용으로 정확도가 다소 떨어졌지만 F2 score값이 약간 상승하는 반비례적인 경향을 보였다. 마지막으로, LR 알고리즘 기반 모델들은 오버샘플링 기법의 효과가 거의 없었다. 위 3개의 알고리즘들 중에서 GBM 알고리즘을 사용하는 모델들이 다른 알고리즘들보다 F2 score값이 전반적으로 높음을 확인하였다. 이 현상은 오버샘플링 기법 적용과 관계없이 동일하게 발생하고 있다.

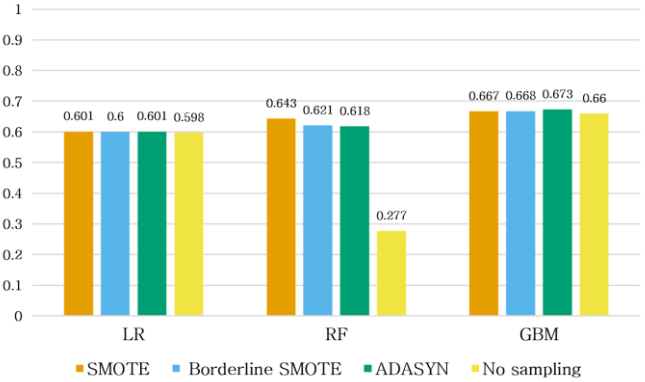

그림 3. 은행 예금 가입 의사 분류 예측 모델들의 F2 score 비교

본 연구에서 제안하는 은행 예금 가입 의사 분류 예측 모델의 목표는 실제로 예금 가입 의사가 있는 고객들을 올바르게 예측하여 예금 가입 성공율을 높이는 것이다. 다시 말해서, 재현율값이 높을수록 또는 재현율 가중치가 높은 F2 score값이 높을수록 우수한 예측 모델이라 할 수 있는 것이다. 따라서, 그림 3을 확인하면 데이터 세트에 ADASYN 기법을 적용한 후 GBM 알고리즘 기반 모델을 사용한 F2 score값이 0.673으로 가장 우수하다고 할 수 있다.

Ⅴ. 결론

본 연구에서는 은행 고객 정보들을 바탕으로 예금 가입 의사가 있는 고객을 예측하는 모델을 개발하였다. 포르투갈 은행 고객 데이터를 사용하였고, 이 데이터 세트가 불균형 데이터이므로 Synthetic Minority Over-sampling Technique (SMOTE), Borderline SMOTE 그리고 Adaptive Synthetic Sampling for Imbalanced Learning (ADASYN) 과 같은 오버샘플링 기법을 적용하였다. 예금 가입 의사를 예측하기 위해 Linear Regression (LR), Random Forest (RF) 및 Gradient Boost Machine (GBM) 등 3 개의 인공지능 알고리즘을 기반으로 한 모델들을 제시하였다. 본 연구의 목표는 실제로 은행 예금을 가입할 의사를 가진 고객들을 올바르게 예측하는 모델 개발인데 모델 성능 지표중의 하나인 재현율이 높은 모델이 이에 해당한다. 다시 말해서, 재현율의 가중치가 높은 F2 score값이 높은 모델이 우수한 성능을 가지고 있는 것이다. 본 연구의 분석 결과로서 ADASYN 샘플링 기법과 GBM 알고리즘을 적용한 모델이 가장 높은 F2 score값을 보였다.

향후 연구에서는 정확도와 F2 score값들에 대한 성능을 향상시키기 위해 새로운 특성들을 추가하거나 오버샘플링 기법의 적용 범위를 넓히는 방법을 모색하고자 한다.

ACKNOWLEDGMENT

Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

참 고 문 헌

[1] Yüksel Akay Ünvan and Ibrahim Nandom Yakubu, “Do Bank-Specific Factors Drive Bank Deposits in Ghana?,” Journal of Computational and Applied Mathematics, vol. 376, no. 112827, Oct. 2020.

[2] Lokesh Lokesh and Iqbal Thonse Hawaldar, “Impact of Factors on the Utilization of Agricultural Credit of Banks: An Analysis from the Borrowers’ Perspective,” Banks and Bank Systems, vol. 14, no. 1, pp. 181- 192, Nov. 2019.

[3] Bank Term Deposit Predictions, “Predictions Subscription to Term Deposits through Marketing Campaigns,” Nov. 2023, Available:(https://www.kaggle.com/datasets/thedevastator/bank-term-depositpredictions/data).

[4] Sipu Hou, Zongzhen Cai, Jiming Wu, Hongwei Du, and Peng Xie, “Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription,” International Journal of Business Analytics, vol. 9, no. 1, pp. 1-12, 2022.

[5] Marcus Hjertaas, Henrik Krantz Knudsen, Jakob Lindstrøm, and Joakim Sælemyr, “Identifying the Best Machine Learning Model for Predicting Bank Term Deposits: An Empirical Study Using Public, Post Financial Crisis Data,” NTNU Handelshoyskolen, 2023.

[6] Much Aziz Muslim, Dasril Yosza, Alamsyah Andi, and Tanzilal Mustaqim, “Bank Predictions for Prospective Long-Term Deposit Investors Using Machine Learning LightGBM and SMOTE,” Journal of Physics: Conference Series, vol. 1918, pp. 1-7, 2021.