JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea. Inst. Inf. Commun. Eng.) Vol. 23, No. 1: 399~406, Mar. 2019

그래디언트 부스팅 모델을 활용한 상점 매출 예측 모델

최재영[1] · 양희윤[2] · 오하영[3*]

Store Sales Prediction Model Using Gradient Boosting Model

Jae-Young Choi[1] · Hee-Yoon Yang[2*] · Hayoung Oh [3*]

[1]Undergraduate Student, Libarary and Information Science, Sungkyunkwan University, Seoul, 03063 Korea
[2]Undergraduate Student, Libarary and Information Science, Sungkyunkwan University, Seoul, 03063 Korea
[3]*Assistant Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

요 약

최근 머신러닝과 딥러닝의 발전에 따라 일상생활과 산업에서 기술을 적용하는 사례들이 많아지고 있다. 금융 및 소비 데이터와 머신러닝 기법을 활용한 연구 또한 활발하게 이루어지고 있다. 본 논문은 이러한 동향에 따라 상점 매출 데이터에 머신러닝 기법을 접목하여 매출 예측 모델을 구축, 핀테크 산업에서의 활용 방안을 제시한다. 다양한 결측치 처리 기법을 적용하고 그래디언트 부스팅 기반의 머신러닝 기법인 XGBoost, LightGBM, CatBoost를 사용하여 각 모델의 상점 매출 예측 성능을 비교한다. 연구 결과, 단일대체법 중 중앙값 대체법을 사용한 데이터셋에 XGBoost를 활용하여 예측을 진행한 모델의 성능이 가장 우수했다. 연구를 통해 얻은 모델을 이용하여 상점의 매출 예측을 진행함으로서 핀테크 기업의 고객 상점들은 대출금을 상환하기 전 금융 보조를 받는 근거로, 핀테크 기업은 상환 가능성이 높은 우수 상점에 금융 상품을 제공하는 등 기업과 고객 모두에게 긍정적인 방향으로 활용할 수 있다.

ABSTRACT

Recently due to the development in machine learning and deep learning, applications of these technologies have been widely utilized daily and industrially. Implementations of machine learning on finance data have been in interest as well. Herein, we employ machine learning algorithms onto store sales data and present future applications for Fintech industries. We consider various missing data processing methods and utilize gradient boosting related machine learning algorithms; XGBoost, LightGBM, CatBoost to predict the future sales for individual stores. As a result, we found that using simple median imputation with a XGBoost model had the best accuracy. By employing the proposed method, stores which have low credibility but have high probability to compensate for repayment can benefit by receiving assistance beforehand, while Fintech enterprises can benefit by offering financial instruments to these stores.

키워드 : 기계 학습, 매출 예측, XGBoost, LightGBM, CatBoost

Keywords : Machine learning, Sales Prediction, XGBoost, LightGBM, CatBoost

Received 29 January 2019, **Revised** 29 March 2019, **Accepted** 21 April 2019
(출판사에서작성)

*** Corresponding Author:** Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://dx.doi.org/10.6109/jkiice.2019.23.1.399 pISSN:2234-4772

Ⅰ. 서 론

1.1 선행연구

머신러닝과 딥러닝이 발전함에 따라 일상생활과 산업에서 이 기술들을 적용하는 사례들이 많아지고 있다. 인터넷 정보검색, 컴퓨터 시각, 음성인식과 언어처리, 모바일 HCI, 생물정보, 바이오메트릭스, 컴퓨터 그래픽, 로보틱스 등 다양한 산업에서 적용되고 있으며 금융계 또한 금융(Finance)과 기술(Technology)을 접목한 핀테크(FinTech) 산업이 발전하고 있다.

머신러닝 기법을 활용한 금융 분야 또한 많은 연구들이 진행되고 있다. 이에 따라, 본 논문에서는 상점 매출 데이터와 머신러닝 기법을 이용한 매출 예측 모델을 제안하고, 핀테크 산업에서의 활용 방안을 제시한다. 현재, 금융 분야의 데이터 분석은 신용카드 연체 예측 모형의 개발[1] 과 신용카드 부도 위험 예측[2] 등의 분류(Classification) 분야와 매출 예측[3]과 주가 예측[4]의 회귀(Regression)분야에서 이루어지고 있다. 분류를 목적으로 하는 연구 중 [1] 연구는 신용 카드사 회원들의 카드 사용 행태와 입금 실적 데이터를 이용해 연체 가능성을 예측했다. 이때 클래스들 간의 구분 경계선을 최적으로 분류하는 SVM(Support Vector Machine)을 통해 회원들의 회원 등급이 정상 혹은 연체 상태일지에 대한 이진 분류를 진행했다. SVM 단일 모델 이외에도 앙상블(Ensemble) 모델링을 시도해 단일 모델보다 앙상블 모델의 정확도가 더 높은 것을 밝혀냈다. 기존의 앙상블 모델은 다양한 머신러닝 기법 모델을 결합하는 방식으로 이루어지나 해당 연구에서는 동일한 SVM 기법을 사용하되, 다양한 데이터 전처리 기법을 통해 생성한 데이터셋으로 여러 모델을 만들어 결합하는 방식으로 앙상블이 이루어졌다. 예측에 사용한 데이터셋이 신용카드의 연체 여부를 포함하고 있어 데이터 불균형(Data Imbalance)을 지니고 있으며 독립변수가 다양한 특성을 갖고 있다. 데이터에 비교적 덜 민감한 부스팅 계열의 LightGBM 등 다른 머신러닝 기법들을 사용한다면 더 높은 정확도가 예상된다. 또 다른 분류 연구인 [2] 의 연구에서는 Kaggle의 UCI Credit Card Dataset[5]을 이용해 머신러닝과 딥러닝 기법 간의 성능을 비교했다. 회귀를 통해 범주에 속할 확률을 0과 1 사이의 값으로 예측하는 로지스틱 회귀(Logistic Regression), SVM(Support Vector Machine), 다수의 의사결정트리(Decision Tree)를 생성하여 앙상블하는 랜덤 포레스트(Random Forest)와 같은 머신러닝 모델과 은닉층(hidden layer)이 두개 이상인 DNN(Deep Neural Network), 합성곱 연산(Convolution) 과 Pooling을 복합적으로 적용한 CNN(Convolutional Neural Network)의 딥러닝 모델 간의 정확도를 비교했으며, 랜덤 포레스트의 정확도가 가장 높다는 결론을 도출했다. 이를 통해 비정형 데이터(이미지, 소리, 문자 등)에서 성능이 뛰어난 DNN, CNN 기법들은 금융 데이터의 정형 데이터에서 머신러닝 기법과 비교해 월등히 높은 정확도를 보이지는 않는다는 것을 알 수 있다. 앙상블 기반 머신러닝 모델인 랜덤 포레스트의 정확도가 높게 나온 것을 바탕으로 더 다양한 방식으로 트리 기반의 앙상블 기법을 적용한다면 성능 향상이 이루어질 수 있을 것이다. 수원시 지방세 체납자 분석 사업의 일환으로 진행된 연구[6] 에서는 수원 시민의 지방세 납부, 체납 정보와 함께 신용 데이터를 결합해 체납자의 자발적 납부 가능성을 사전에 예측해 이를 체납여부를 분류하는 것을 목표로 한다. 체납된 세금의 회수 가능성 예측 모형을 개발하기 위해서 랜덤 포레스트, 로지스틱 회귀, 재귀적으로 변수의 분할을 실시하는 반복분할(Recursive Partitioning), 신경망(Neural Network) 등의 다양한 모델을 사용했다. 결과적으로 로지스틱 회귀 방식을 채택했다. 로지스틱 회귀 방식은 통계적으로 해석 가능하다는 점에서 공무 집행의 확실한 근거가 되어줄 수 있으나, 데이터가 축적되어 규모가 커지면 보다 효율적이고 정확한 예측이 가능한 모델이 성능적으로 좋을 것으로 예상된다. 회귀를 목적으로 하는 연구인 [3]의 연구에서는 Kaggle 의 Rossmann Store Sales 데이터[7]를 이용해 시계열 데이터의 머신러닝 예측을 진행했다. 규제항을 추가하여 과적합을 방지하는 라쏘 회귀모형(Lasso Regression)과 랜덤 포레스트, 랜덤포레스트에 무작위성을 부여한 Extra Tree, 자기회귀와 이동평균을 고려하는 ARIMA(Autoregressive Integrated Moving Average), 신경망 모델과 Extra Tree, 그리고 XGBoost의 여러 다른 모델들을 쌓아 만드는 스태킹(Stacking) 모델을 이용해 예측을 시도했다. 결과적으로, 스태킹 모델의 정확도가 가장 좋았으며 시계열 데이터에서의 트리 계열 모델이 우수함을 증명했다. 또 다른 회귀 연구인 [4]의 연구에서는 인공신경망 알고리즘인 CNN과 순환신경망의 RNN(Recurrent Neural Network)을 이용해 KOSPI 상위 20위 종목의 주가를 예측했으며, 정확도는 RNN이 52%로 가장 높았다. [8]의 연구는 양방향의 순환신경망인

LSTM에 attention 메커니즘을 결합한 Attention-Bi-LSTM 기법을 통해 무선 이동통신사 KT, LG유플러스, SK텔레콤의 KOSPI 주가를 예측했다. 이 때 Attention Layer의 Context Vector를 이용하는 방식에 따라 세 가지 방법으로 예측을 진행했으며, 양방향 LSTM Hidden Layer의 마지막 시퀀스 Vector와 Attention Layer의 Context Vector을 합성곱(Element-Wise Dot Product)한 방안의 정확도가 가장 좋았다. 두 연구([4],[8]) 모두, 다양한 신경망 모델을 사용하여 주가를 예측하는 데에 의의가 있으나 머신러닝 기법과 모델에 대한 스태킹, 앙상블을 진행한다면 성능 향상과 함께 해석력을 갖출 수 있을 것으로 예상된다.

1.2 연구 동향

이처럼, 금융 데이터에서는 기본적인 머신러닝 기법인 로지스틱 회귀, 랜덤 포레스트, 신경망 모델(DNN, RNN, CNN)이 사용되고 있지만, 트리 기반 앙상블 모델들은 사용된 경우는 드물다. XGBoost, LightGBM, CatBoost 등의 그래디언트 부스팅 기반 모델들은 초미세먼지 예측[9] 등의 환경 분야, 아파트 실거래가 예측[10] 등의 부동산 분야 그리고 전력 수요 예측[11] 과 안전 운전자 예측[12] 등의 공학 분야에서 사용되고 있다. 표 1을 통해 선행 연구를 확인할 수 있다.

이러한 연구 동향에 따라, 본 논문에서는 상점의 과거 매출 기록 데이터[13]를 바탕으로 머신러닝 기법을 이용해 미래의 매출액을 예측하는 방안을 분석했다. 선행 금융 데이터 연구에서 주로 사용되지 않았던 그래디언트 부스팅(Gradient Boosting) 기반의 머신러닝 기법인 XGBoost, LightGBM, CatBoost를 사용하고, 데이터 전처리 과정에서 여러 결측치 기법을 적용해 각 모델의 성능을 비교했다.

본 논문의 구성은 다음과 같다. 먼저, 제 II 장에서 s 논문에서 사용된 결측치 처리 알고리즘과 머신러닝 알고리즘을 살펴본다. 제 III 장은 연구 대상과 데이터 전처리 과정을 설명하고, 제 IV장에서 머신러닝 모델들을 이용한 예측을 진행하고, 그 결과에 대해 검토한다. 마지막으로 제 V 장에서는 결론과 시사점을 도출한다.

Table.1 Tasks of Past Studies

Task	Goal	Model
Classification	Credit Card Delinquency Prediction[1]	SVM, Ensemble of SVM
Classification	Default Payments of Credit Card Clients in Taiwan[2]	Logistic Regression, SVM, Random Forest, DNN, CNN
Classification	Suwon Citizen Delinquency Prediction[6]	Random Forest, Logistic Regression, Recursive Partitioning, DNN
Regression	Rossmann Store Sales Prediction[3]	[Stacking] Lasso Regression, ARIMA, Extra Tree, XGBoost, DNN
Regression	KOSPI Stock Price Prediction[4]	DNN, CNN, RNN
Regression	KT,LG U+,SKT Stock Price Prediction [8]	Attention-Bi-LSTM
Classification	PM2.5 in Seoul Prediction[9]	XGBoost, Ensemble of XGBoost
Regression	Seoul Residential Apartment Price Prediction[10]	[Stacking] XGBoost, LightGBM
Regression	Load Forecasting[11]	XGBoost
Regression	Safety Driver Preidicition[12]	XGBoost, LightGBM

Ⅱ. 분석 알고리즘

2.1 결측치 처리 알고리즘

2.1.1 결측치 메커니즘

결측치를 처리하기 위해서는 결측치가 발생하는 메커니즘을 고려해야 한다. 각 변수에서의 결측값과 관측된 값의 관계에 따라 MCAR, MAR, MNAR 3가지 경우로 나눌 수 있다. MCAR(Missing Completely At Random)은 결측치의 발생이 관측값, 결측값과 완전히 무관하여 독립적인 경우이다. 이는 데이터셋의 어떠한 데이터와도 관련 없이 결측이 발생하기 때문에 결측치의 양이 적은 경우에 대체 및 삭제를 진행해도 데이터 전체의 변동성에 적은 영향을 미친다. MAR (Missing At Random)은 결측값 이외의 관측값만이 결측치의 발생과 관련이 있는 경우로, 관측값들로 결측값을 유추해 낼 수 있으며 대부분의 통계적 대체 방법론들이 이를 전제로 한다. MNAR(Missing Not At Random)는 결측치가 관측값과 결측값 모두와 관련이 있는 경우다.[14] 각각의 메커니즘