JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 2: 171~177, Feb. 2021

그래디언트 부스팅 모델을 활용한 상점 매출 예측

최재영[1] · 양희윤[1] · 오하영[2*]

Store Sales Prediction Using Gradient Boosting Model

Jaeyoung Choi[1] · Heeyoon Yang[1] · Hayoung Oh[2*]

[1]Undergraduate Student, Library and Information Science, Sungkyunkwan University, Seoul, 03063 South Korea
[2*]Associate Professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 South Korea

요 약

최근 머신러닝의 발전에 따라 일상생활과 산업에서 기술을 적용하는 사례들이 많아지고 있다. 금융 데이터와 머신러닝 기법을 활용한 연구 또한 활발하게 이루어지고 있다. 본 논문은 이러한 동향에 따라 상점 매출 데이터에 머신러닝 기법을 접목해 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 활용해 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

Through the rapid developments in machine learning, there have been diverse utilization approaches not only in industrial fields but also in daily life. Implementations of machine learning on financial data, also have been of interest. Herein, we employ machine learning algorithms to store sales data and present future applications for fintech enterprises. We utilize diverse missing data processing methods to handle missing data and apply gradient boosting machine learning algorithms; XGBoost, LightGBM, CatBoost to predict the future revenue of individual stores. As a result, we found that using median imputation onto missing data with the appliance of the xgboost algorithm has the best accuracy. By employing the proposed method, fintech enterprises and customers can attain benefits. Stores can benefit by receiving financial assistance beforehand from fintech companies, while these corporations can benefit by offering financial support to these stores with low risk.

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

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

Received 9 November 2020, **Revised** 5 December 2020, **Accepted** 19 December 2020

*** Corresponding Author** Hayoung Oh (E-mail:hyoh79@gmail.com, Tel:+82-2-583-8585)
Associate Professor, College of Computing & Informatics, Sungkyunkwan University, Seoul, 03063 South Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.2.171 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

1.1. 선행연구

머신러닝이 발전함에 따라 정보검색, 모바일 HCI, 바이오메트릭스 등 다양한 산업에서 적용되고 있고, 금융계 또한 금융(Finance)과 기술(Technology)을 접목한 핀테크(Fin-Tech) 산업이 발전하고 있다. 이러한 동향에 맞춰, 머신러닝 기법을 활용한 금융 분야 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 크게 분류와 회귀 분석 두 분야로 이루어지 있다. 분류를 목적으로 하는 연구 중 [1] 의 연구에서는 Kaggle의 UCI Credit Card Dataset[2]을 이용해 머신러닝과 딥러닝 기법 간의 성능을 비교했다. 회귀를 통해 범주에 속할 확률을 0과 1 사이의 값으로 예측하는 로지스틱 회귀(Logistic Regression), 클래스들 간의 구분 경계선을 최적으로 분류하는 SVM(Support Vector Machine), 다수의 의사결정트리(Decision Tree)를 생성하여 앙상블하는 랜덤 포레스트(Random Forest)와 은닉층이 두 개 이상인 DNN(Deep Neural Network), 그리고 합성곱 연산(Convolution) 과 Pooling을 복합적으로 적용한 CNN(Convolutional Neural Network)의 정확도를 비교한 결과, 랜덤 포레스트의 정확도가 가장 높다는 결론을 도출했다. 또 다른 분류 연구인 [3] 에서는 신용카드 사용기록을 통해 부정 이용 여부를 예측하는 분류 모델을 개발하는 것을 목표로 했다. 로지스틱 회귀, 의사결정트리 그리고 신경망 모델(Neural Network) 을 이용해 신용카드 부정사용의 분류를 진행했으며, 결과적으로 로지스틱 회귀와 신경망 모델이 높은 정확도를 기록했다. 회귀연구인 [4]의 연구에서는 Kaggle 의 Rossmann Store Sales 데이터[5]를 이용해 시계열 데이터의 머신러닝 예측을 진행했다. 규제항을 추가하여 과적합을 방지하는 라쏘 회귀 모형(Lasso Regression)과 랜덤 포레스트, 랜덤포레스트에 무작위성을 부여한 Extra Tree, 자기회귀와 이동평균을 고려하는 ARIMA(Auto-regressive Integrated Moving Average), 신경망 모델과 Extra Tree, 그리고 XGBoost를 쌓아 만드는 스태킹(Stacking) 모델을 이용해 예측을 시도했다. 결과적으로 스태킹 모델의 정확도가 가장 높아 시계열 데이터에서 트리 계열 모델이 우수함을 증명했다. 또 다른 회귀 연구인 [6]에서는 인공신경망 알고리즘인 CNN과 순환신경망의 RNN(Recurrent Neural Network)을 이용해 KOSPI 상위 20위 종목의 주가를 예측했으며, 정확도는 RNN이 52%로 가장 높았다.

1.2. 연구 동향

이처럼, 금융 데이터에서는 다양한 머신러닝 기법인 로지스틱 회귀, 랜덤 포레스트, 신경망 모델(DNN, RNN, CNN)이 사용되고 있지만, 트리 기반의 앙상블 모델 사용은 드물다. XGBoost, LightGBM, CatBoost 등의 그래디언트 부스팅 기반 트리 앙상블 모델들은 집값 예측[7] 등의 부동산 분야, 초미세먼지 예측[8] 등의 환경 분야, 그리고 전력 수요 예측[9] 과 안전 운전자 예측[10] 등의 공학 분야에서 사용되고 있다. [7] 의 연구에서는 XGBoost, CatBoost, 랜덤포레스트 등의 머신러닝 기법을 이용해 플로리다의 집값을 예측했다. 다양한 머신러닝 기법 중 XGBoost의 성능이 가장 좋았다. 또한, 연속형인 종속변수(집값)를 범주화해 학습시킨 후, 그 결과를 다시 연속형 변수에 학습, 최종적으로 집값을 예측하는 Target Binning 방식에서도 XGBoost와 CatBoost 의 정확도가 높아, 그래디언트 부스팅 기반의 트리 모델이 회귀예측에 좋은 성능을 보이는 것을 확인할 수 있다.

이러한 연구 동향에 따라, 본 논문에서는 상점의 과거 매출 기록 데이터[11]를 바탕으로 선행 금융 데이터연구에서 보편적으로 사용되지 않았던 그래디언트 부스팅(Gradient Boosting) 기반의 머신러닝 기법인 XGBoost, LightGBM, CatBoost를 사용해, 미래의 매출액 예측에 대해 연구했다. 또한, 데이터 전처리 과정에서 다양한 결측치 기법을 적용해 각 모델의 성능을 비교했다.

본 논문의 제 II 장에서는 논문에서 사용된 연구 대상과 연구 방법을 설명하고, 제 III 장은 머신러닝 모델들을 활용한 실험결과를 도출한다. 마지막 제 IV 장에서는 결론과 시사점을 제안한다.

Ⅱ. 연구 방법

본 논문은 상점의 신용카드 매출 데이터에 다양한 결측치 알고리즘을 적용해 앙상블 기반의 트리계열 머신러닝 모델로 예측 성능을 비교 및 평가하는 것을 목표로 한다. 세부 예측 목표는 과거의 매출 데이터를 이용해

이후 3개월 매출을 예측하는 것이다. 이를 위해 국내 데이터 경쟁 플랫폼 DACON[12]에서 진행된 '상점 신용카드 매출 예측 경진대회[11]의 데이터를 사용했으며 세부적인 변수는 표1과 같이 구성되어있다. 전처리 과정 중 결측치 처리 부분은 평균 대체법, 중앙값 대체법, 선형 보간법, 곡선 보간법[13-15] 등의 기법들을 사용했다. 예측 모델은 그래디언트 부스팅[16] 계열의 XGBoost[17-18], LightGBM[19-20], CatBoost [21-22] 모델을 사용했다.

Table. 1 Train Set Variables[11]

store_id	Store ID
card_id	Credit Card ID
card_company	De-identified Card Company Name
transacted_date	Date of Transaction
transacted_time	Time of Transaction
installment_term	Installment Term of Transaction
region	Store Region
type_of_business	Type of Business
amount	Transaction Amount

결측치가 각각 31%, 60%를 차지하는 region 변수와 type_of_business 변수를 제거했고, 상점의 거래액을 예측하는 데 영향을 미치지 않을 것으로 판단한 card_id, card_company, transacted_time, installment _term 변수를 제거했다. 학습 데이터의 거래액이 0인 행을 삭제했고, 음수의 경우에는 환불이 이루어진 것으로 판단해 store_id와 card_id 가 동일한 양수 거래액을 찾아 함께 제거했다.

예측 목표가 각 상점의 미래 3개월 매출이므로, 거래가 이루어진 시간으로 기록 되어 있던 기존 데이터 셋을 각 상점의 월 매출 합계 데이터로 만들었다.

월별 매출의 특성을 강조하기 위해, 3개월 전 ~ 12개월 전의 각 상점의 월별 매출을 갖는 three_month,⋯ twelve_month를 추가했다. 이 변수는 과거의 값을 이용했기 때문에, 해당 시점으로부터 n개월 이전의 데이터가 없다면 결측값을 가진다. 이를 처리하는 방식에 따라 2가지 방향으로 데이터셋을 생성했다. NA 데이터셋의 경우, 3개월 전 ~ 12개월 전 변수 생성 과정(Date Shifting 과정)에서 발생하는 결측값을 유지해 완성했다. 결측치 처리 데이터셋의 경우, 변수 생성 과정에서 발생한 매출의 결측값을 평균 대체법, 중앙값 대체법, 선형 보간법, 곡선 보간법으로 결측치를 처리해 결과적으로 MEAN, MEDIAN, LINEAR, SPLINE 데이터셋을 생성했다. 전처리 과정을 통한 최종 데이터셋의 변수는 표 2와 같다.

학습데이터의 경우, 상점별 매출기록의 시작시점과 종료 시점이 동일하지 않은 것을 고려하여 예측 기간을 2018년 12월 ~ 2019년 2월까지의 3개월로 고정하지 않고 각 상점 별로 매출이 기록된 마지막 3개월로 설정했다. 그림 1과 같이, 772번 상점은 2017년 10월 ~ 2018년 2월의 매출 기록이 부재하고, 2018년 10월에 기록이 종료된다. 이 경우에는, 학습 데이터는 2017년 6월 ~ 2018년 7월, 검증 데이터는 2018년 8월 ~ 2018년 10월의 매출기록이 되는 것이다.

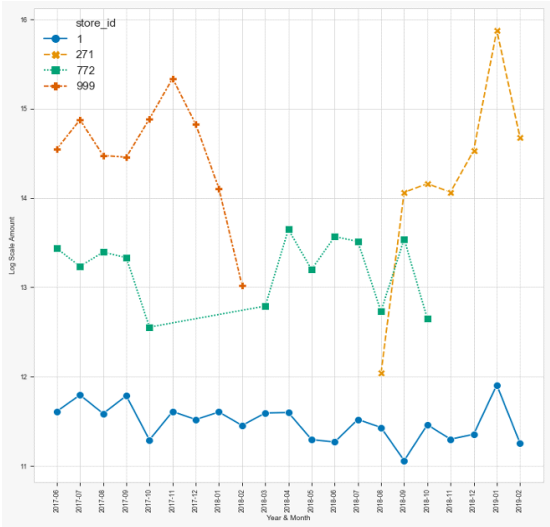

Fig. 1 Difference Between Time Series Plot for Store ID 1, 271, 772, 999

Table. 2 Final Features for the Train Data

store_id	Store ID
month	Month of Transaction Date
year	Year of Transaction Date
three_month	Total Transaction Amount "nth" Month before the Transaction Month
four_month	
five_month	
six_month	
seven_month	
eight_month	
nine_month	

ten_month	
eleven_month	
twelve month	
amount	Total Transaction Amount for the “month”

Ⅲ. 실험 및 결과

3.1. 모델 학습 데이터 및 평가 지표

모델 학습에 사용한 데이터셋은 총 5개로, 결측치가 보존된 NA 데이터셋과 평균 대체법, 중앙값 대체법, 선형 보간법, 곡선 보간법을 적용한 MEAN, MEDIAN, LINEAR, SPLINE 4개의 데이터셋이다. 과거의 데이터를 이용해 3개월 이후의 매출을 예측하되 매출 기록이 상점별로 연속적이지 않다는 점을 고려해, 상점별로 기록된 매출의 마지막 3개월 데이터를 예측했다.

모델 학습의 결과를 평가하는 지표로는 MAE(Mean Absolute Error)를 사용했다. 관측값과 예측값 사이의 차이를 절댓값으로 나타낸 것을 모두 더해 평균을 낸 것으로 다음의 수식 (1)로 나타낼 수 있다. 여기서 y_t는 참값, $\hat{y_t}$는 예측값이다.

$$MAE = \frac{\sum |y_t - \hat{y_t}|}{n} \quad (1)$$

3.2. 모델 학습

앞서 구축한 5개의 데이터셋을 이용해 각 데이터셋에 XGBoost, LightGBM, CatBoost, 총 3개의 모델로 예측을 진행했다. 과적합을 방지하기 위해 교차검증을 시행하되, 상점별로 매출을 예측해야 한다는 점을 고려해 상점들이 각 Fold에 고르게 분포할 수 있도록 store_id를 기준으로 Stratified 5-Fold를 학습과정에서 적용했다. 결측치를 포함한 NA 데이터셋을 기준으로 Random Grid Search 와 Bayesian Optimization 기법을 혼용해 각 모델의 하이퍼파라미터를 최적화했다.

3.3. 모델 검증

모델 학습을 통해 도출한 최적의 하이퍼파라미터들을 이용해 검증한 결과는 표 3와 같다.

XGBoost의 경우 결측치를 중앙값으로 대체한 MEDIAN 데이터셋이 성능이 가장 좋았으며, SPLINE 데이터셋의 성능이 가장 좋지 않았다. LightGBM의 경우, NA 데이터셋, LINEAR 데이터셋, SPLINE 데이터셋, MEDIAN 데이터셋, MEAN 데이터셋 순서로 정확도가 낮아지는 것을 볼 수 있다. CatBoost은 MEAN 데이터셋에서 가장 높은 정확도를 보이며, 두 번 째로 높은 정확도를 보인 데이터셋은 MEDIAN 데이터셋이다.

Table. 3 MAE Test Scores Results

	XGBoost	LightGBM	CatBoost
NA	321279.78	320953.13	360355.44
MEAN	323330.41	326836.75	353011.75
MEDIAN	320487.36	325040.74	353409.89
LINEAR	323462.43	323223.69	359235.33
SPLINE	327696.05	323269.50	354742.81

15개의 모델-데이터셋 조합을 검증한 결과, 각 모델을 기준으로 가장 좋은 성능을 보이는 조합은 MEDIAN-XGBoost 데이터셋, NA- LightGBM 데이터셋, MEAN-CatBoost 데이터셋이다. 이에 따라, 결측값을 각 상점 매출의 중앙값으로 대체한 데이터셋을 XGBoost로 학습한 모델과 결측값이 보존된 데이터셋을 LightGBM을 통해 학습한 모델, 그리고 결측값을 각 상점 매출의 평균값으로 대체한 데이터셋을 CatBoost로 학습한 모델을 최종 모델로 선정했다.

Table. 4 Comparisons of MAE and Time Measure for the Final Models

Dataset	Model	MAE	Time(s)
Median	XGBoost	320487.36	3537.50
NA	LightGBM	320953.13	9.70
Mean	CatBoost	353011.75	135.70

표 4의 경우, 선정된 데이터셋-모델 각각의 MAE와 이를 학습하는데 소요되는 시간(초)을 보여준다. MAE 값이 가장 낮은 모델은 중앙값 대체 데이터셋을 활용한 XGBoost이며, 근소한 차이로 결측값 보존 데이터셋을 학습한 LightGBM이 두 번째로 낮다. 그러나 시간 소요의 측면에서 LightGBM이 단 9초 만에 검증 데이터셋을 예측하기 때문에 XGBoost는 LightGBM에 비해 효율적이지 못하다고 추론할 수 있다.

그림 2는 무작위로 선정된 10개의 상점의 매출 예측값을 시각화한 그래프이다. 실제 값(actual amount)는 엑스 선으로 표현했고, 중앙값 대체 데이터셋과 XGBoost

의 예측값은 세모선, 결측치 데이터셋과 LightGBM의 예측값은 네모선, 마지막으로, 평균 대체 데이터셋과 CatBoost의 예측값은 다이아몬드 선으로 나타냈다. 대체적으로 중앙값 대체 데이터셋을 이용한 XGBoost 모델의 예측값이 실제값과 유사한 것을 확인할 수 있다.

Ⅳ. 결 론

본 논문은 결측치 알고리즘과 그래디언트 부스팅 알고리즘을 적용해 상점별 매출 예측을 진행했다. 이를 통해 결측치 알고리즘과 머신러닝 기법에 따른 예측 정확도를 비교할 수 있었다. 상점 매출 데이터에 평균 대체법, 중앙값 대체법, 선형 보간법, 곡선 보간법을 적용해 총 5가지 데이터셋을 생성했다. 이 5가지 데이터셋과 XGBoost, LightGBM, CatBoost 모델을 이용해 예측을 진행한 결과, 각각 차례로 중앙값 대체 기법을 이용한 데이터셋, 결측치가 보존된 데이터셋, 평균값 대체기법을 이용한 데이터셋과의 조합에서 좋은 성능을 보였으며 그 중 중앙값 대체 기법과 XGBoost를 사용한 모델의 MAE 값이 가장 낮아 최종 모델로 선정했다.

각 모델에서 우수한 성능을 보인 데이터셋이 결측치 대체 알고리즘 중 단일대체법을 적용한 것으로 보아, 대체법보다 복잡한 결측치 보간법 알고리즘을 사용해 결측치를 대체하는 것이 항상 좋은 성능을 내지 않는다고 추론 할 수 있다. 또한, XGBoost의 예측 정확도가 다른 모델의 예측 정확도 보다 우수하다는 점에서 성능이 좋다고 평가할 수 있으나, 속도 측면에서 유사한 성능을 보이는 LightBGM보다 시간 효율적이지 못하다는 점에서 독보적이라고 할 수 없다.

본 논문에서는 상점 매출 데이터의 변수가 다양하지 않아 분석에 한계가 있었다. 또한, 그래디언트 부스팅 기반의 XGBoost, LightGBM, CatBoost 만을 활용해 예측을 진행했으나 향후 연구에서는 ARIMA 등의 시계열 기법, DNN, LSTM 등의 딥러닝 기법 그리고 여러 기법들을 함께 사용하는 스태킹 기법을 활용한다면 상점의

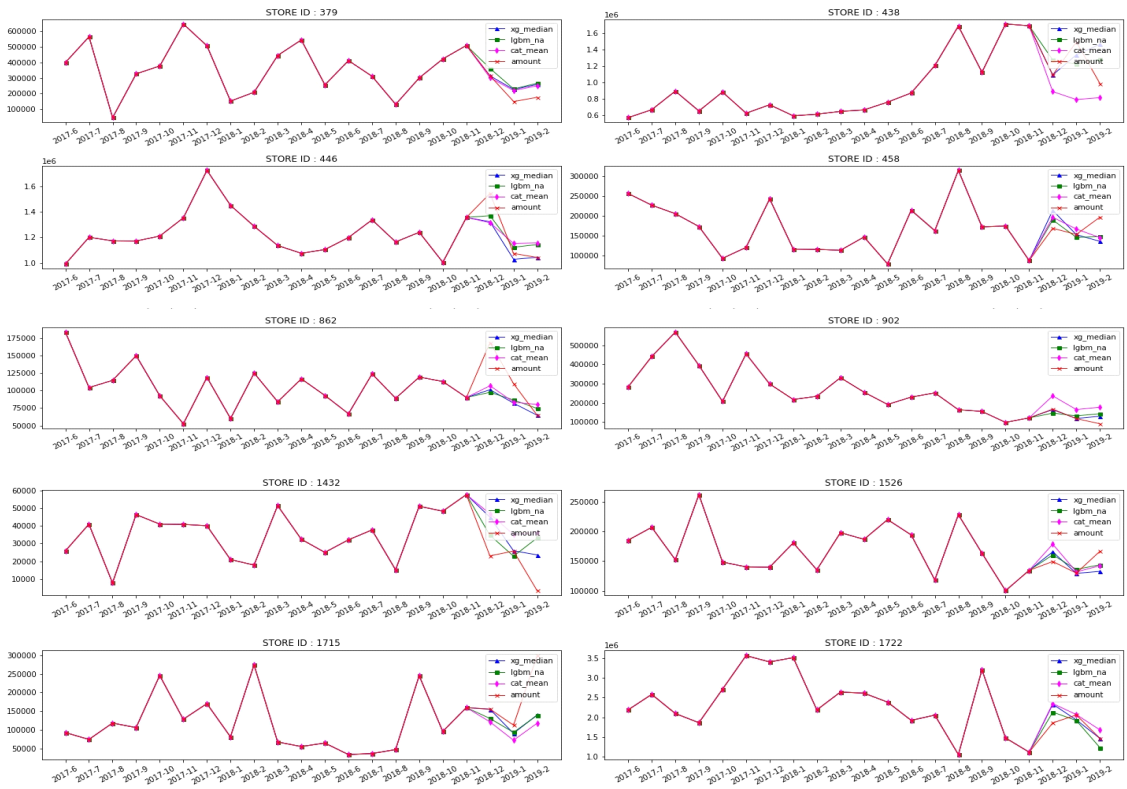

Fig. 2 Comparisons of MAE Time for Final Models

특성을 더 반영하며 추가적인 성능 향상을 기대할 수 있을 것이다. 본 논문에서 제시한 매출 예측 모델을 통해 상인들을 대상으로 대출 서비스를 운영하는 핀테크 기업들에게 상인들을 위한 맞춤 서비스를 진행하는데 도움을 줄 수 있을 것으로 예상한다. 신용 점수가 낮거나 담보를 갖지 못하지만, 높은 매출액이 예상되어 상환이 잘 이루어질 것으로 보이는 우수 상점들에게는 미리 금융 서비스를 제공하는 등의 방식으로 예측 모델을 활용할 수 있을 것으로 전망한다.

REFERENCES

[1] J. M. Yoon, "Effectiveness Analysis of Credit Card Default Risk with Deep Learning Neural Network," *Journal of Money & Finance*, vol. 33, no. 1, pp. 151-183, Mar. 2019.

[2] Kaggle. UCI Credit Card Dataset [Internet]. Available: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.

[3] A. Shen, R. Tong, and Y. Deng, "Application of Classification Models on Credit Card Fraud Detection," in *2007 International Conference on Service Systems and Service Management*, pp. 1-4, Jul. 2007.

[4] B. M. Pavlyshenko, "Machine-Learning Models for Sales Time Series Forecasting," *Data*, vol. 4, no. 1, Apr. 2019.

[5] Kaggle. Rossmann Store Sales [Internet]. Available: https://www.kaggle.com/c/rossmann-store-sales.

[6] J. H. Lee, "Stock price prediction model using deep learning," M. S. Thesis, Soongsil University, Seoul, 2016.

[7] S. B. Jha, R. F. Babiceanu, V. Pandey, and R. K. Jha, "Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study," *arXiv: 2006.10092v1*, Jun. 2020.

[8] H. Kim, "The Prediction of PM2.5 in Seoul through XGBoost Ensemble," *Journal of the Korean Data Analysis Society*, vol. 22, no. 4, pp. 1661-1671, Aug. 2020.

[9] Y. G. Lee, J. Y. Oh, and G. B. Kim, "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, Feb. 2020.

[10] S. I. Jang and K. C. Kwak, "Comparison of Safety Driver Prediction Performance with XGBoost and LightGBM," in *Proceeding of Korea Institute of Infomation Technology Conference*, pp. 360-362, Jun. 2019.

[11] Dacon. Korea data competition platform. Card Sales Prediction contest [Internet]. Available: https://dacon.io/competitions/official/140472/overview/.

[12] Dacon. Korea data competition platform [Internet]. Available: https://dacon.io/.

[13] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. Hambrug, NJ: John Wiley & Sons Inc., 2014.

[14] S. R. Lee, "Comparison of algorithms for the missing data imputation methods," M. S. Thesis, Hankuk University of Foreign Studies, Seoul, 2020.

[15] Yonsei Structure & Bridge Eng Lab. Interpolation [Internet]. Available: http://str.yonsei.ac.kr/korean/portal.php.

[16] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 2nd ed. Cambridge, MA: The MIT Press., vol. 29, no. 5, pp.1189-1194, 2001.

[17] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco: CA, pp. 785-794, 2016.

[18] Documents for Xgboost [Internet]. Available: https://xgboost.readthedocs.io/en/latest/#.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach: CA, pp. 3149-3157, 2017.

[20] Documents for Lightgbm [Internet]. Available: https://lightgbm.readthedocs.io/en/latest/index.html.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, V. A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical feature," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 6639-6649, 2018.

[22] Documents for Catboost [Internet]. Available: https://catboost.ai/.

최재영(Jaeyoung Choi)

성균관대학교 문헌정보학과
※ 관심분야 : 머신러닝, 딥러닝, NLP

양희윤(Heeyoon Yang)

성균관대학교 문헌정보학과
※ 관심분야 : 머신러닝, 딥러닝, 컴퓨터비전, 추천시스템

오하영(Hayoung Oh)

성균관대학교 소프트웨어융합대학 글로벌융합학부 부교수
서울대학교 컴퓨터 공학 박사
이화여자대학교 컴퓨터 공학 석사
※ 관심분야 : 소셜정보망 및 데이터 분석