영화 및 드라마 컨텐츠의 온라인 리뷰 빅데이터 기반 시청자 의견과 감성 트랜드 텍스트 분석

임보민a, 김경원a, 신수민b,*∗*

*a국립 인천대학교 글로벌정경대학 무역학부*

*b국립 한국교통대학교 사회복지학과*

# Abstract

*Background:* Korea has the highest suicide rate among Organisation for Economic Co- operation and Development (OECD) countries. Consequently, central and local governments and private organizations in Korea cooperate in promoting various suicide prevention projects to actively respond to suicide problems. Machine learning has been used to predict suicidal ideation in the fields of health and medicine but not from a social science perspective.

*Objective:* Since suicidal ideation is a major predictor of suicide attempts, being able to

anticipate and mitigate it helps prevent suicide. Therefore, this study presents a data-based analysis method for predicting suicidal thoughts quickly and effectively and suggests countermeasures against the causes of suicidal thoughts.

*Participants and Methods:* To predict early signs of suicidal ideation in children and adolescents, big data collected for approximately 4 years (from 2017 to 2020) from the Korea Youth Policy Institute (NYPI) were used. To accurately predict suicidal ideation, supervised ma- chine learning classification algorithms such as logistic regression, random forest, XGBoost, multilayer perceptron (MLP), and convolutional neural network (CNN) were used.

*Results:* Using CNN, suicidal ideation was predicted with an accuracy of approximately

90%. The logistic regression results showed that sadness and depression increased suicidal thoughts by more than 25 times, and anxiety, loneliness, and experience of abusive language increased suicidal thoughts by more than three times.

*Conclusions:* Machine learning and deep learning approaches have the potential to predict

and respond to suicidal thoughts in children, adolescents, and the general population, as well as help respond to the suicide crisis by preemptively identifying the cause.

*Keywords:* Demand Forecasting, Explainable Prediction, Machine and Deep Learning, SHAP, Business Decision

*∗*Corresponding author: thekimk.kr@gmail.com

*ORCID(s):* 0000-0001-6530-8426 (K. Kim)

# Introduction

[글로벌 고속철도의 발전과 중요성]

[한국의 고속철도 이용현황으로 중요성 설명]

[고속철도 수요예측 기존연구에서 정성적 또는 정량적 연구 사례 소개 및 한계]

[머신러닝 및 딥러닝으로 수요예측을 하려는 기존연구들의 장점과 단점 소개 + 최신 인공지능 알고리즘 사용한 수요예측 필요성 강조]

[설명가능한 인공지능 소개 및 XAI 활용 수요예측으로 비즈니스 의사결정 기여가능 강조]

[본 연구의 목적과 기여 요약]

1. 실제 비즈니스에 활용가능한 미래 예측 성능 달성
2. 예측된 미래수요 기반 XAI 활용 정책적 의사결정 지원
3. 고성능PC 없이 현장에서 쉽게 활용가능한 XAI 활용 고성능 비즈니스 애널리틱스 방법론 제시

[논문의 구성]

# Methods

* 1. *Participants and Data Preprocessing*

그림 2. 데이터 모델링을 위한 전처리 과정

* 1. *Machine Learning Algorithm: Bagging and Boosting*

기본적으로 머신러닝 알고리즘은 예측 오차를 줄이는 방향으로 설계되었다. 오차는 bias and variance로 분리될 수 있는데, 예측의 안정성에 초점을 두어 variance를 줄이기 위해 샘플링 기법을 활용하는 Bagging과 성능에 초점을 두어 bias를 줄이기 위해 반복적인 모델링을 활용하는 Boosting으로 구분될 수 있다. Bagging은 여러 개의 샘플을 추출하여 각 샘플마다 모델링한 후 outputs의 투표로 최종 output의 label을 결정하는 방식이기에 bootstrap aggregating으로 불린다. 본 연구에 활용된 대표적 알고리즘은 Random Forest(RF)다. RF는 samples를 병렬처리로 한꺼번에 모델링 할 수 있기 때문에 속도도 빠르다며 training 데이터의 학습 성능이 우수한 편이고 noise data에도 성능이 크게 변하지 않는 robust algorithm이다 (Michie, Spiegelhalter, and Taylor 1999). 변수의 중요도를 제공하긴 하지만 각 샘플별 모델링 과정에서 각 변수들의 우선순위를 평균한 것으로 positive or negative와 같은 영향력의 방향성을 포함하지는 못하는 단점이 있다.

Bagging처럼 Boosting도 classification and regression 문제 모두 활용할 수 있는 supervised learning이다 (Chen and Guestrin 2016). Gradient boosting methods(GBM)를 기반으로 Extreme gradient boosting (XGBoost) and LightGBM, CatBoost 등의 알고리즘으로 확장되며 단점들이 개선되고 다양한 기능이 추가되었다. 본 연구에서는 XGBoost, LightGBM and CatBoost를 대표적인 boosting 알고리즘으로 활용하여 자하철 혼잡도 예측에 사용한다. 이 알고리즘들은 전체 데이터를 학습할 때 잘 학습되지 않은 에러들을 가중치를 높여 재학습을 하면서 성능을 향상시키는 방향으로 설계되었다. 메모리를 효율적으로 사용하거나 computation 속도를 높이기 위한 많은 기능들이 추가되었고, iterative learning과 병렬처리 과정에서 성능이 높아질 수 있었다 (Alsubari et al. 2021).



그림 3. 배깅(Bagging)과 부스팅(Boosting) 알고리즘의 구조 비교 (Cha, Moon, and Kim 2021).

* 1. *Deep Learning Algorithm: MLP, RNN*

딥러닝은 인공지능의 한 방법론으로, 인간의 두뇌 구조에서 영감을 받아 개발되었다. 데이터의 복잡한 패턴들을 학습하기 위해 연속된 layer를 중첩하여 변수들의 모든 상호작용을 포함하여 의미있는 규칙들을 학습해 내는데 강점이 있다. 이러한 구조의 기본이 되는 알고리즘으로 multilayer perceptron (MLP)가 있으며, 이미지나 시계열 등의 데이터도 학습해 낼 수 있도록 layer의 구조나 흐름을 개선하여 convolutional neural network (CNN)과 recurrent neural network (RNN) 등으로 확장되었다. neural network structure를 가진 모든 알고리즘들은 앞서 소개한 머신러닝의 알고리즘 처럼 activation function만 선택적으로 변경함으로써 classification and regression 문제 모두에 활용가능한 universal approximator이다. 모델의 구조가 복잡하기에 수많은 가중치들을 한꺼번에 추정해야 하는데 feed-forward and backpropagation 을 사용하여 loss function을 최소화하기 위해 gradient descent optimizer를 사용하였다 (see Figure 2) (Viswavandya, Patel, and Sahoo 2021). 머신러닝과 마찬가지로 다양한 nonlinear patterns를 학습함으로써 기존의 알고리즘보다 성능이 대폭 향상되어 인간이 인지하기 어려운 것들도 학습해내는데 유용하다.



그림 4. 대표적인 딥러닝 알고리즘인 MLP의 구조와 데이터 학습을 통한 수많은 가중치 추정 과정

본 연구에서는 MLP and CNN 알고리즘을 활용하여 지하철 혼잡도를 예측한다. 특히 CNN은 이미지 분류 특화 알고리즘이지만 내부 구조가 결국 feature를 요약하여 것이기 때문에 regression 문제에도 활용이 가능하다. Feature를 요약하기 위해 convolution, pooling, and fully connected layers를 활용하여 인접한 feature values들의 관련성에 더욱 초점을 두어 학습한다. 이러한 과정에서 여러 개의 neurons로 구성된 layer들을 계속적으로 통과하는데, 결국 low-level feature vector에서 점차 high-level feature vector로 변수들의 특성이 변환되어 학습된다 (Shustanov and Yakimov 2017). 따라서 다양한 변수들의 상호작용을 포함한 특징들이 학습될 것이고 마지막 output은 지하철 혼잡도 값이 출력되며 regression 문제를 해결한다.

* 1. *MLP 기반 알고리즘*
  2. *RNN 기반 알고리즘*
  3. *Model Explainability: SHapley Additive exPlanations (SHAP)*

앞서 소개한 인공지능의 대표적인 알고리즘인 머신러닝과 딥러닝은 발생가능한 변수들의 모든 상호작용들을 스스로 생성하여 학습하기 때문에 인간의 수준을 뛰어넘는 성능 달성이 가능하게 하였다. 하지만 왜 그러한 결과가 발생한 것인지 쉽게 확인하거나 설명하기는 어려운 블랙박스 알고리즘이다. 일부 머신러닝 알고리즘이 변수들의 중요도(Feature Importance)를 출력하지만 방향성이 없기 때문에 해석에 주의를 기울여야 한다. 하지만 SHAP(Shapley Additive exPlanations)는 local interpretable model-agnostic explanations (LIME)과 shapley value를 연결한 이론으로, 이러한 블랙박스와 같은 알고리즘들의 한계를 보완하여 지하철 혼잡도 예측에 영향을 주는 변수들의 특징과 방향을 설명해준다. LIME은 주어진 데이터 값들을 변화시킬 때 모델 예측값의 변화와의 관련성을 가중치로 계산한다 (Molnar 2023). 그리고 shapley values는 coalitional game theory를 기반으로 변수들의 기여도를 계산하는 metric으로 생각할 수 있다. 따라서 변수들이 가질 수 있는 모든 coalitions 경우들을 생성한 후 실게 값들이 입력되었을 때 변화된 예측값의 기여도 평균을 계산한다. 하지만 연산량이 매우 많을 수 있기 때문에 랜덤 샘플링 기법을 활용 계산의 효율성을 높인다. 이 2가지 이론이 결합된 SHAP를 사용하여 지하철 혼잡도에 기여하는 변수들의 정도와 방향성을 알 수 있고 블랙박스와 같은 인공지능 알고리즘을 설명가능케 한다.

선형 기반의 단순한 기울기로 해석하는 전통적인 statistical algorithms는 우리가 직관적으로 이해하기는 쉽지만 샘플이 1개라도 변경되면 해석이 얼마든지 달라질 수 잇는 불안정성이 있고 실제 예측 성능도 낮아서 단순히 설명만 될 뿐 실제 비즈니스에 활용하기는 어렵고 조심해야 한다. 반면, 인공지능 알고리즘은 구조가 복잡하고 정교하지만 높은 정확성을 나타낸다. 따라서 인공지능 알고리즘을 사용하되 SHAP와 같은 설명가능한 기능들을 결합하며 설명력을 강화하기 위한 model explainability의 중요성은 점점 높아지고 있다.

* 1. *Evaluation Metrics*

지하철 혼잡도를 예측하고 설명의 신뢰성을 높이기 위해 6개의 regression metrics를 사용하여 모델링의 성능을 확인한다. 이들은 RMSE(Root Mean Squared Error), MSPE(Mean Squared Percentage Error), MAE(Mean Absolute Error), MAPE(Mean Absolute Percentage Error), MedAE(Median Absolute Error), MedAPE(Median Absolute Percentage Error)이다. 이러한 six evaluation metrics는 아래와 같이 계산된다:

모든 metrics는 예측 성능이 좋을수록 낮은 수치들이 나오도록 실제값()과 예측값()의 차이로 구성되어 있다. 따라서 6개의 metrics 모두 낮은 수치가 나오는 경우 지하철 혼잡도를 잘 예측하는 모델이라 볼 수 있다.

# Results

# Conclusions

# References

Alsubari, Saleh Nagi, Sachin N. Deshmukh, Mosleh Hmoud Al-Adhaileh, Fawaz Waselalla Alsaade, and Theyazn H. H. Aldhyani. 2021. 'Development of Integrated Neural Network Model for Identification of Fake Reviews in E-Commerce Using Multidomain Datasets', *Applied Bionics and Biomechanics*, 2021: 5522574.

Cha, Gi-Wook, Hyeun-Jun Moon, and Young-Chan Kim. 2021. 'Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables', *International journal of environmental research and public health*, 18: 8530.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. San Francisco, California, USA: Association for Computing Machinery.

Michie, D., D. Spiegelhalter, and Charles Taylor. 1999. 'Machine Learning, Neural and Statistical Classification', *Technometrics*, 37.

Molnar, Christoph. 2023. *Interpretable machine learning. A Guide for Making Black Box Models Explainable* (Lulu. com).

Shustanov, Alexander, and Pavel Yakimov. 2017. 'CNN Design for Real-Time Traffic Sign Recognition', *Procedia Engineering*, 201: 718-25.

Viswavandya, Meera, Shashwat Patel, and Kaushik Sahoo. 2021. 'ANALYSIS AND COMPARISON OF MACHINE LEARNING APPROACHES FOR TRANSMISSION LINE FAULT PREDICTION IN POWER SYSTEMS', *Journal of Research in Engineering and Applied Sciences*, 6: 24-31.