

# TeGCN:씬파일러 신용평가를 위한 트랜스포머 임베딩 기반 그래프 신경망 구조 개발

김성수\*  
연세대학교 정보대학원  
(brady\_kim@yonsei.ac.kr)

배준호\*  
연세대학교 정보대학원  
(juunho@yonsei.ac.kr)

이주현\*  
연세대학교 정보대학원  
(jay248@yonsei.ac.kr)

정희주\*  
하나금융IT 클라우드센터 MSP셀  
(heejoo@hanafn.com)

김희웅  
연세대학교 정보대학원  
(kimhw@yonsei.ac.kr)

국내 씬파일러(Thin Filer)의 수가 1200만명을 넘어서며, 금융 업계에서 씬파일러의 신용을 정확히 평가하여 우량고객을 선별해 대출을 공급하는 시도가 많아지고 있다. 특히, 차주의 신용정보에 존재하는 비선형성을 반영하여 채무불이행을 예측하기 위해서 다양한 머신러닝 알고리즘을 활용한 연구가 진행되고 있다. 그 중 그래프 신경망 구조(Graph Neural Network)는 일반적인 신용정보 외에 대출자 간의 네트워크 정보를 반영할 수 있다는 점에서 데이터가 부족한 씬파일러의 채무 불이행 예측에서 주목할 만하다. 그러나, 그래프 신경망을 활용한 기존의 연구들은 신용정보에 존재하는 다양한 범주형 변수를 적절히 처리하지 못했다는 한계가 있었다. 이에 본 연구는 범주형 변수의 맥락적 정보를 추출할 수 있는 트랜스포머 메커니즘(Transformer mechanism)과 대출자 간 네트워크 정보를 반영할 수 있는 그래프 합성곱 신경망(Graph Convolutional Network)를 결합하여 효과적으로 씬파일러의 채무 불이행 예측이 가능한 TeGCN (Transformer embedded Graph Convolutional Network)를 제안한다. TeGCN은 일반 대출자 데이터셋과 씬파일러 데이터셋에 대하여 모두 베이스라인 모델 대비 높은 성능을 보였으며, 특히 씬파일러 채무 불이행 예측에 우수한 성능을 달성했다. 본 연구는 범주형 변수가 많은 신용정보와 데이터가 부족한 씬파일러의 특성에 적합한 모델 구조를 결합하여 높은 채무 불이행 예측 성능을 달성했다는 시사점이 있다. 이는 씬파일러의 금융소외문제를 해결하고 금융업계에서 씬파일러를 대상으로 추가적인 수익을 창출하는데 기여할 수 있을 것이다.

**주제어** : 씬파일러, 채무 불이행 예측, 그래프 합성곱 신경망, 범주형 변수 임베딩, TeGCN

논문접수일 : 2023년 9월 6일  
원고유형 : Regular Track

논문수정일 : 2023년 9월 22일  
교신저자 : 김희웅

게재확정일 : 2023년 9월 24일

## 1. 서론

국내 씬파일러(Thin Filer)의 수가 1200만명을 넘어서며, 신용등급에 따른 금융 양극화가 심각해지고 있다(김나경, 김예지, 2023; 김정산, 2023). 씬파일러는 주로 취업준비생, 사회초년생, 노년층,

주부 등 금융거래 실적이 적은 ‘금융이력부족자’를 의미한다(기획재정부, 2021). 이들은 대출, 신용카드 등 상환능력을 평가할 수 있는 거래 이력의 정보가 부족하기 때문에 일반적으로 신용점수가 낮게 평가된다(권영탁, 2021). 실제로 2022년 말 기준 국내 개인신용평가사 코리아크레딧뷰로(KCB)

\* These authors contributed equally to this work

의 신용평점에 따르면 찢파일러의 99.9%는 800점 미만의 중·저신용자에 해당했다. 이들은 과소평가된 신용점수로 인해 금융서비스에서 소외되고 있다.

최근 들어 찢파일러의 채무 이행 능력을 정확히 평가하여 신규 고객으로 확보하려는 기업들이 많아지고 있다(김하영 등, 2022). 먼저, 이는 찢파일러가 갖춘 경제적 성장가능성에 기인한다(네이버 테크핀 리포트, 2021). 찢파일러의 46%는 2030세대인 것(김연정, 2021)으로 나타났는데, 이들은 향후 소득 수준이 증가할 것으로 예상된다(통계청, 2021). 이에 찢파일러 고객을 미래 장기 고객으로 유치하고자 하는 것이다. 또한, 찢파일러 대출상품에 대한 인센티브가 증가하는 점 역시 찢파일러에 대한 관심이 높아지는 이유이다. 금융당국은 2022년 가계대출 총량 규제에서 중금리 대출을 제외했다(김유진, 2022). 이로 인해 은행들은 빠르게 성장할 것으로 기대되는 찢파일러 대출 시장을 잡기 위해 찢파일러 고객의 신용을 정확히 측정할 수 있는 신용평가모형에 많은 투자를 하고 있다.

이에 머신러닝, 딥러닝 알고리즘을 활용하여 채무 불이행 예측모델을 개발하는 연구들이 활발하게 진행되고 있다(엄하늘 등, 2020). 일례로, 이근희 등(2017)은 합성곱 신경망(Convolutional Neural Network)를 활용한 딥러닝 신용평가모형을 개발하였다. 또한 오윤철(2020)은 K-평균 군집화, XGBoost, Random Forest 등의 모델을 활용한 P2P 대출상환 여부 예측모델을 개발하였다. 이우형(2018)은 확률기반 부스팅을 활용하여 중·저신용자에 특화된 신용평가모형을 제안하였으며, 로지스틱 회귀를 활용하여 신용점수를 도출하는 모델 역시 개발되기도 하였다(Woo and Sohn, 2022).

최근에는 대출 신청자 간의 관계를 반영한 네트워크 정보를 활용하여 신용정보의 부족을 해소

하고, 정확하게 채무 불이행을 예측하는 연구들이 주목받고 있다(Calabrese et al., 2017; Agosto et al., 2019). 대표적으로, 대출 신청자의 가상 공간 관계를 활용하여 네트워크 구조를 생성한 연구(Lee et al., 2021), 그래프 합성곱 신경망(Graph Convolutional Network, 이하 GCN)을 활용하여 신용평가 시 발생하는 신용정보 결측치 문제를 완화한 연구(Li et al., 2022; Lee et al., 2021)가 존재한다. 그러나, 대출 신청자의 네트워크 정보를 활용하는 기존 연구들은 신용평가 모델 학습을 위해 효과적인 범주형 변수 처리를 고려하지 못했다는 한계가 존재한다. 일반적으로 개인신용 정보는 연속형 변수와 범주형 변수가 혼합되어 있는 형태를 띤다(유성준, 박나리, 2020). 하지만 기존 연구는 주로 범주형 변수의 처리를 유클리드 공간의 벡터로 처리하기 위해 원-핫 인코딩을 수행하였다. 원-핫 인코딩은 매우 높은 차원의 벡터 표현으로 인해 모델의 학습 성능이 떨어지고 컴퓨팅 비용이 증가하는 문제를 야기할 수 있다(Vong et al., 2019; Seger, 2018). 따라서 다양한 범주형 변수를 포함하고 있는 신용 데이터를 적절히 표현하고, 신용 정보가 부족한 찢파일러의 특징을 반영할 수 있는 차별화된 신용평가모델 구조가 필요하다.

이에 본 연구는 범주형 변수의 맥락을 반영한 임베딩이 가능한 Tab Transformer 구조와 찢파일러의 네트워크 정보를 반영한 TeGCN (Transformer-embedded Graph Convolutional Network) 모델을 제안한다. 기존의 머신러닝/딥러닝 기반 채무 불이행 예측모델과 달리, TeGCN 모델은 신용정보에 존재하는 다양한 범주형 변수를 효과적으로 처리할 수 있으며, 찢파일러의 데이터 부족 문제를 차주 네트워크를 통해 극복할 수 있다. 이를 통해 찢파일러의 채무 불이행을 정확히 예측하고, 금융 소외 문제 해결에 기여할 수 있다.

## 2. 개념적 배경

채무 불이행 예측은 대출자 또는 대출 예정자가 앞으로 채무를 이행할 가능성을 예측하여, 그 가능성이 높은 그룹과 낮은 그룹으로 분류하는 작업을 의미한다(Zhou & Wang, 2012). 채무 불이행 예측은 은행 등 금융사들의 가장 중요한 관심사 중 하나이다(Abraham & Zhang, 2008). 채무 불이행은 은행의 자본 적정성(capital adequacy)에 문제를 야기하고, 심할 경우 банкр던까지 이어질 수 있기 때문이다(Li et al., 2022). 이에 은행들은 다양한 방법론을 활용하여 채무 불이행을 예측하고자 노력해왔다(Baensens et al. 2003). 그러나 전통적인 통계기반 모델은 경제, 금융, 신용 변수 간의 비선형적인 관계를 측정하는데 적절하지 않은 것으로 드러났다(Zhang et al., 2023; 김명중, 2012). 따라서, 최근에는 복잡한 비선형 관계를 다루기 위한 머신러닝, 딥러닝 기반의 채무 불이행 예측 연구가 주목을 받고 있다(Luo et al., 2017; Zhou et al., 2020).

### 2.1 기계학습 기반 채무 불이행 예측

머신러닝, 딥러닝 기반 모델은 우수한 결과를 도출하기에 신용평가기관에서 고객 대출금액 산출이나 위험예측 등에 활발하게 활용되고 있다(김성진, 안현철, 2016). 먼저 He and Fan (2021)은 앙상블 기법을 활용하여 채무 불이행을 예측하는 모델을 개발했다. 이때, LightGBM과 합성곱 신경망을 앙상블하여 변수 표현(feature representation)을 학습한 뒤 Inner Product-based Neural Network(IPNN)을 분류기로 활용했다. 이외에도 다양한 샘플링 기법과 머신러닝 알고리즘을 결합하여 채무 불이행 예측 모델을 개발한 사례 역시 존재한다. Moscato et al.

(2021)은 랜덤 언더 샘플링, 랜덤 오버 샘플링, SMOTE 샘플링 등 서로 다른 샘플링 기법을 활용하여 로지스틱 회귀, 랜덤포레스트, MLP(multi-Layer Perception) 모델을 각각 학습시켰다. Fu et al. (2020)은 P2P 플랫폼 내의 리뷰에 주목하였다. 이 연구는 P2P 플랫폼 이용자들이 작성하는 리뷰에서 BiLSTM 모델을 활용하여 주요 키워드를 추출하였다. 이후 추출한 키워드를 바탕으로 감성점수 등 변수들을 생성하여 채무 불이행을 예측했다. 위 연구에서 제안한 다양한 모델은 기존 머신러닝 모델 대비 분류 성능 기준으로 더 나은 성능을 달성했다.

이렇듯 딥러닝 및 머신러닝을 활용한 채무 불이행 예측에 대한 연구가 지속적으로 진행되고 있음에도 불구하고, 여전히 그 범위는 제한적이다. 이는 과거 대출 데이터의 부족과 대출 신청자의 불완전한 프로필 및 누락된 정보 처리의 어려움 때문에 발생하는 정확성의 한계 때문이다(Li et al., 2022). 이러한 한계점을 극복하기 위해 채무자 간 관계성을 추가적으로 활용하는 연구들이 증가하고 있다(Lee and Sohn, 2021). 본 연구는 데이터 간 관계성을 효과적으로 모델링할 수 있는 그래프 신경망(Graph Neural Network)를 활용하여 채무자 간 관계 네트워크를 형성하고, 이를 바탕으로 데이터 누락문제를 해결하고자 한다.

### 2.2 그래프 합성곱 신경망 기반 채무 불이행 예측

네트워크 데이터는 복잡하고 비구조적인 특성을 갖기에 정형 데이터에서 사용되는 일반적인 모델 구조로는 처리가 제한적이다. 이에 네트워크에서 정보를 효과적으로 학습할 수 있는 모델 구조가 필요하다(Muñoz-Cancino et al., 2023). 네트워크

데이터를 처리하기 위해 등장한 그래프 신경망은 데이터 개체 간의 연결을 그래프 형태로 모델링하여 이웃 정보의 특징을 효과적으로 포착할 수 있으며, 네트워크 기반 변수에서 고차원적인 (higher-order) 이웃 관계 분석을 통해 숨겨진 패턴을 반영할 수 있다는 장점이 있다(Lagasio et al., 2022). 그 중, 합성곱 연산을 활용하는 GCN은 그래프 형태의 데이터에서 각 노드의 고유한 특징과 노드 간의 관계를 이용해 분류 문제를 해결하기 위해 제안된 딥러닝 기법이다(Kipf & Welling, 2016). GCN은 노드 간의 복잡한 관계를 효과적으로 반영함으로써, 네트워크 내의 잠재된 패턴을 파악하고 분석할 수 있다(Lee et al., 2021; Li et al., 2022). 따라서, 해당 구조를 통해 대출 신청자 간의 관계성 정보를 학습하고 채무 불이행을 예측하는 연구들이 증가하고 있다(Li et al., 2022, Lee et al., 2021).

먼저, Shumovskaia et al. (2021)은 대출 신청자 간의 연결 강도와 고객 간 이체의 규칙성 사이에 밀접한 연관성이 있다는 점에 주목하였다. 이를 바탕으로, 그들은 대출 고객 네트워크를 분석하기 위한 GCN 기반의 노드 간 링크 예측 모델을 제안하였다. 해당 모형은 채무 불이행 예측을 직

접으로 수행하는 방법론은 아니었으나, 그래프 구조를 활용함으로써 신용 평가의 정확성을 향상시킬 수 있는 가능성을 보여주었다.

Lee et al. (2021)은 대출 신청자 간의 관계를 중요한 정보 중 하나로 파악하고, GCN 기반의 채무 불이행 예측 모델을 개발하였다. 이 연구에서는 대출 신청자의 정보를 대출 정보, 신용 이력 정보, 그리고 차주 간 관계성 정보 세 가지 카테고리로 세분화하였다. 이를 위해, 해당 정보 유형을 효과적으로 반영할 수 있는 세 개의 그래프 모델을 독립적으로 구성하였으며, 이를 통합하는 방식으로 대출 신청자의 복합적인 특성을 포착하였다.

Li et al. (2022)은 유사한 대출 신청 기록이 유사한 채무 불이행 위험으로 연결될 가능성에 집중하였다. 이에 무방향 그래프 기반의 GCN 모델을 제시하였으며, 해당 모델은 유사한 이웃 노드 간의 연결을 포착하여 채무 불이행 예측에 있어 누락된 값을 보완하는 부가 정보로 활용하였다. GCN 등 머신러닝/딥러닝 방법론을 활용하여 채무 불이행 예측 모델을 개발한 선행연구들은 표 1과 같다.

그러나 기존 연구에서는 그래프 네트워크를 이용하여 대출 신청자 간의 연결 정보를 계산하는 과정에서 범주형 변수를 적절히 처리하지 못한

〈표 1〉 머신러닝/딥러닝을 활용한 채무 불이행 예측 연구

저자	방법론	활용 데이터
He & Fan (2021)	LightGBM, CNN	DefaultData dataset, PPDai dataset
Moscato et al. (2021)	Logistic Regression, MLP, Random Forest	Lending Club
Fu et al. (2020)	BiLSTM	Wangdaizhijia (중국 P2P 대출 플랫폼)
Lee and Sohn (2021)	Spatial Probit Model	Lending Club
Shumovskaia et al. (2021)	GCN	Fund Transfers Dataset
Lee et al. (2021)	GCN	Lending Club
Li et al. (2022)	GCN	Home Credit Default Risk dataset, Lending Club

한계점이 있다(Lee et al., 2021). 수치형 변수 간의 유사도는 유클리드 거리 등으로 비교적 쉽게 계산할 수 있으나, 범주형 변수는 이러한 방식의 유사도 계산이 불가능하다. 이로 인해, 대출자 간의 관계를 정확하게 측정하는데 있어 어려움이 존재하며, 대출자 정보의 범주형 변수를 효과적으로 반영할 수 있는 새로운 접근 방식이 필요하다.

### 2.3 범주형 변수 임베딩

머신러닝 알고리즘들은 수치형 변수만을 입력값으로 고려하기 때문에, 범주형 변수들을 수치형으로 변환하는 것이 필수적이다(Dahouda & Joe, 2021). 이에 많은 연구에서는 원-핫 인코딩(one-hot encoding)과 같은 방법을 통해 문자열 값을 수치형으로 변환하여 범주형 변수를 다루려고 하였다. 그러나 원-핫 인코딩된 변수는 필요 이상으로 고차원적이며, 변수 간의 내재적 유사성을 적절히 반영하지 못하는 어려움이 존재한다(Cerda et al., 2018; Chollet, 2021).

이를 극복하기 위하여, 인공 신경망 구조를 활용하여 범주형 변수의 표현력을 높이는 연구들이 진행되고 있다. 먼저 Guo and Berkahn (2016)은 범주형 변수를 다차원 공간의 벡터 형태로 변환하여 변수의 표현력을 높일 수 있는 엔티티 임베딩(Entity Embedding) 방법론을 제안하였다. 원-핫 인코딩된 독립변수와 종속변수를 활용하여 인공 신경망을 학습시키고, 학습과정에서 도출한 가중치를 입력값을 의미하는 벡터값으로 활용했다. 또한, Duarte et al. (2018)은 양방향 게이트 순환 유닛(BiGRU)과 어텐션 메커니즘을 활용한 범주형 변수 임베딩 방법론을 제시했다. 특히, Huang et al. (2020)은 각 범주형 변수의 각 범주(class)가 독립적으로 표현되는 MLP 기반 임베딩의 한계에

주목하여, 범주형 변수의 맥락 정보를 반영할 수 있는 column embedding과 Transformer 메커니즘을 활용한 TabTransformer를 제안했다. 해당 연구에서 제안한 TabTransformer를 활용한 범주형 변수 임베딩은 MLP 기반 임베딩 대비 뛰어난 성능을 보였으며, 결측치나 이상치가 많은 데이터에 대해서도 높은 성능을 달성했다.

개인신용정보는 연속형 변수와 범주형 변수가 혼합되어 있는 형태를 띤다(유성준, 박나리, 2020). 따라서 신용정보 데이터 내의 범주형 변수를 적절히 표현하지 못할 경우 모델의 성능이 떨어질 수 있다(Seger, 2018). 특히 씬파일러 신용정보 데이터의 경우 결측치의 비율이 높기에 변수 내 표현력을 높이는 것이 매우 중요하다. 이에 본 연구는 범주형 변수 내의 맥락 정보의 반영이 가능하며, 결측치가 많은 경우에도 강건한 성능을 보이는 TabTransformer 구조를 활용하여 변수의 표현력을 개선하고, 궁극적으로 씬파일러의 채무 불이행 예측 성능을 높이고자 한다.

## 3. 연구 방법론

### 3.1 모델 설계

본 연구는 범주형 변수 및 결측치 비율이 높은 변수 표현에 강점을 가진 TabTransformer 구조(Huang et al., 2020)와 데이터 개체 간의 관계성을 효과적으로 포착할 수 있는 GCN 구조(Zhang et al., 2019)를 결합하여 씬파일러의 채무 불이행을 정확히 예측할 수 있는 TeGCN을 제안한다. 이를 위하여, 먼저 신용정보의 범주형 변수와 연속형 변수를 구분하여 TabTransformer에 텐서로 입력한다. 이후 임베딩된 각 변수의 벡터를 결합

(Concatenate)한 뒤 GCN레이어에 입력한다. 이때 GCN 레이어의 각 노드(Node)는 개별 차주의 신용정보를 뜻하며, 각 엣지(Edge)는 차주 간의 관계 강도를 의미한다. 이를 통해 이웃 노드의 관계정보가 반영된 GCN 레이어의 출력값은 최종적으로 완전연결층(Fully Connected Layer)과 시그모이드 함수를 거쳐 채무 불이행 예측값을 반환하게 된다. 본 모델의 구조는 그림 1과 같다.

### 3.2 TabTransformer: 범주형 변수 임베딩

TabTransformer는 범주형 변수의 맥락적 정보를 반영하도록 임베딩하여 예측 정확도를 높이기 위해 제안된 딥러닝 기법이다. Column embedding 레이어, N개의 Transformer 레이어, 다층 퍼셉트론 레이어로 구성되어 있으며, 각 Transformer 레이어는 multi-head self-attention 레이어와 position-wise feed-forward 레이어로 구성되어 있다.

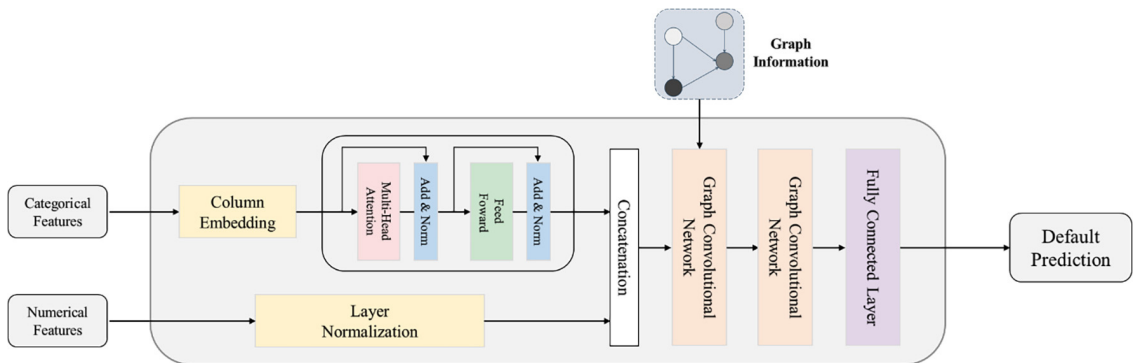
TabTransformer는 범주형 변수와 연속형 변수에 대하여 각각 다른 임베딩 방법론을 적용한다. 먼저 범주형 변수는 column embedding 과 Transformer 레이어를 통과하여 오차를 계산하기 위한 다층 퍼셉트론 레이어로 입력된다. 이때 column embedding

이란 같은 범주형 변수 내의 범주끼리 일정 부분 같은 parameter를 공유하도록 임베딩하는 것을 의미한다. 이를 통해 모델은 각 범주형 변수가 갖는 서로 다른 특성을 학습한다. 또한, 변수 내의 각 범주 간 상관관계가 높은 경우 가까운 유클리디안 거리를 갖도록 임베딩된다.

이를 구체적으로 설명하자면,  $i$  변수의  $j$  범주에 대한  $d$ 의 길이를 갖는 임베딩 벡터  $e_{\varphi_i}(j)$ 는 아래의 식 1과 같이 표현되게 된다.  $c_{\varphi_i}$ 는 범주형 변수의 각 범주끼리 공유하는 파라미터의 벡터를 의미하며,  $w_{\varphi_{ij}}$ 는 해당 값에 대하여 자체적으로 학습하는 파라미터의 벡터를 의미한다.  $c_{\varphi_i}$ 와  $w_{\varphi_{ij}}$ 를 병합(concatenate)하여 도출된  $e_{\varphi_i}(j)$ 는 각 범주형 변수가 갖는 자체적인 특성을 유지함으로써 타 범주형 변수와 구별되게 된다. 이때  $c_{\varphi_i}$ 의 차원  $l$ 은 변수 별 임베딩 시 어느 정도의 비율로 파라미터를 공유할지 선정하는 하이퍼 파라미터에 해당한다.

$$e_{\varphi_i}(j) = [c_{\varphi_i}, w_{\varphi_{ij}}] \quad \text{식 (1)}$$

where  $c_{\varphi_i} \in R^l, w_{\varphi_{ij}} \in R^{d-l}$



〈그림 1〉 TeGCN 모델 구조

이후 column embedding을 거친 임베딩  $E_\phi(x_{cat})$ 은 N개의 Transformer 레이어에 순차적으로 투입된다. 이 과정에서 각 임베딩은 다른 임베딩의 맥락을 반영하여 맥락 임베딩(Contextual Embedding)  $\{h_1, h_2, \dots, h_m\}$ 으로 변환되어 출력된다. 마지막으로, 맥락 임베딩은 연속형 변수  $x_{cont}$ 와 결합되어 다층 퍼셉트론 레이어의 입력값으로 투입되며, 경사하강법에 의해 오차를 줄여가며 각 파라미터를 업데이트하게 된다.

### 3.3 그래프 정보: 대출 신청자 군집화

본 연구에서는 그래프 네트워크를 설계하기 위해 Li et al. (2022)의 방식을 차용하여, 동질적인 범주형 변수를 가진 대출 신청자들을 군집화하였다. 이러한 접근은 여러 차원을 가진 범주형 변수에서 대출 신청자 간의 관계를 보다 명확하게 이해하기 위한 것이며, 그래프 구성 시 사용되는 파라미터를 결정하는데 도움을 줄 수 있다.

전체 데이터셋에서  $i$ 개의 대출 신청자를 갖고 있을 때, 차주 ( $B$ )의 데이터셋  $D$ 는 여러 명의 차주인  $D = \{B_1, B_2, \dots, B_i\}$ 로 구성된다. 각 대출 신청자  $B_i$ 는  $B_i = [x_{i1}, x_{i2}, \dots, x_{ij}]$ 와 같은 신용 정보를 포함하고 있다. 여기서,  $x_{ij}$ 는  $i$ 번째 신청자에 대한  $j$ 번째 변수의 값이며, 각 변수는 수치형 변수와 범주형 변수로 구성되어 있다. 범주형 변수 군집화는 각 범주형 변수가 갖고 있는 고유한 값의 수에 따라 1차적으로 필터링 후 진행된다. 이는 특정 변수들이 너무 많은 차원을 갖고 있을 경우, 대출 신청자 관계를 파악하기 위한 군집화의 수행 목적을 달성하기 어렵기 때문이다. 본 연구에서는 군집화의 기준이 되는 범주형 변수 값의 차원 수를 선정하기 위해 5, 10, 15, 20으로 설정하여 실험을 수행하였고, 차원의 수가 10일

때 가장 높은 성능을 달성하였다.

이를 바탕으로 군집화는  $j$ 개의 변수  $x_{i1}, x_{i2}, \dots, x_{ij}$  중 차원의 개수가 10이하인 범주형 변수에 대하여 수행한다. 예를 들어,  $1, 2, \dots, n$ 의 범주형 변수 중  $1, 3, n$ 번째 변수의 차원이 10 이하일 경우, 클러스터링의 기준이 되는 변수  $\tau$ 는  $x_{i1}, x_{i3}, x_{in}$ 가 된다. 이렇게 선정된 변수  $\tau$ 를 기준으로 그룹화하여 동일한 범주형 변수 값을 갖는 대출 신청자를 하나의 그룹으로 묶는다. 이때, 전체  $n$ 개의 범주형 변수 중 동일한 값을 갖는 모든 경우의 조합을 파악하여, 동일한 조합을 가진 그룹을 군집으로 구성한다. 따라서, 각 대출 신청자  $B_i$  중 클러스터링 변수  $\tau$ 가 동일한 값을 갖는  $B_i$ 는  $k$ 번째 군집  $C_k$ 에 속하게 된다. 이 과정을 통해 전체 대출 신청자 데이터셋  $D$ 에 포함되어 있는 데이터 포인트들은 고유한 ID를 갖는 여러 군집으로 그룹화된다. 그 결과  $i$ 개의 대출 신청자에서  $k$ 개의 클러스터가 구성될 수 있으며, 군집 집합  $C = \{C_1, C_2, \dots, C_k\}$ 이다.

### 3.4 GCN: 그래프 형성

본 연구에서 GCN은 그래프  $G = (V, E)$ 는 노드 집합  $V$ 와 엣지 집합  $E$ 로 구성된다. 대출 신청자를 노드로 간주하면, 관계에 따라 노드들 사이에 가중치가 부여된 엣지를 생성하여 무방향 그래프를 형성할 수 있다.

그림 2는 GCN의 각 레이어에서의 전파(propagation) 규칙을 나타낸다. GCN 모델은 노드의 특징을 표현하는 피쳐 행렬  $H$ 와 그 노드들 간의 관계를 나타내는 인접 행렬  $A$ 가 존재한다. 각 GCN 레이어에서는 피쳐 행렬  $H$ 에 인접 행렬  $A$ 가 곱해진다. 이 과정을 통해 대출 신청자의 노드 특징을 해당 노드의 이웃 노드 특징과 통합할

수 있다. 그 결과로 생성된 출력 행렬은 다음 레이어의 노드 피쳐 행렬로 사용된다.  $K$ 번째 GCN 레이어에서의 노드 피쳐 행렬  $H$ 는 다음의 식 2와 같이 계산될 수 있다.

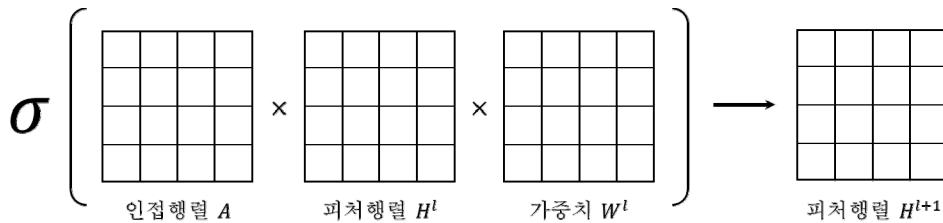
$$H^{(l+1)} = \sigma(A \times H^{(l)} \times W^{(l)} + b^{(l)}) \quad \text{식 (2)}$$

$H^{(l)}$ 은  $l$ -번째 레이어의 노드 피쳐 행렬을,  $W^{(l)}$ 는  $l$ -번째 레이어의 학습 가중치 행렬을 나타낸다. 또한  $b^{(l)}$ 는  $l$ -번째 레이어의 편향(bias)을 표현하며,  $W^{(l)}$ 와  $b^{(l)}$ 의 학습 가능한 파라미터(learnable parameter)들은 공유된다.  $\sigma$ 는 비선형 활성화 함수로써, ReLU와 Sigmoid와 같은 함수들이 사용될 수 있다. 인접 행렬  $A$ 와  $H^{(l)}W^{(l)}$ 의 곱을 통해, 노드 정보가 이웃한 노드에 의해서만

갱신(update)되는 GCN의 특성을 구현하게 된다.

본 연구에서는 대출 신청자 간의 관계를 무방향(undirected) 그래프로 표현한다. 이러한 그래프를 구축하기 위해 입력 벡터의 고유한 값들을 기반으로 인접 행렬(adjacency matrix)과 차수 행렬(degree matrix)을 생성한다. 구체적으로, 입력 벡터  $a$ 의 각 원소는 데이터 포인트가 속한 클러스터의 인덱스를 나타낸다. 본 연구에서는 대출 신청자 간의 관계를 무방향(undirected) 그래프로 표현한다. 이러한 그래프를 구축하기 위해 입력 벡터의 고유한 값들을 기반으로 인접 행렬(adjacency matrix)과 차수 행렬(degree matrix)을 생성한다. 이 과정에 대한 상세한 의사코드는 아래와 같다.

이 과정을 통해 군집 정보를 바탕으로 연결 행렬과 차수 행렬을 도출한다. 이후 GCN 네트워크의



〈그림 2〉 GCN의 전파 규칙

---

**Algorithm:** Matrix Construction

---

**Input:**

*Dataset:* A complete dataset.

*ClusterTensor:* Tensor containing cluster IDs for each row in *Dataset*.

**Output:**

Constructed *adjacency matrix* and *degree matrix*.

**Procedure:**

Initialize an empty dictionary *IndexDict*

**for** each unique value  $\mathcal{V}$  in *ClusterTensor* **do:**

Identify and store indices where the value is  $\mathcal{V}$  in *IndexDict*

Initialize an *adjacency* and a *degree matrix* to zeros of *ClusterTensor* length

**for** each index *IndexCol* and value  $\mathcal{V}$  in *ClusterTensor* **do:**

Establish connections in the *adjacency matrix* based on *IndexDict*[ $\mathcal{V}$ ]

Update diagonal of *degree matrix* based on the size of cluster  $\mathcal{V}$

Calculate the inverse square root of *degree matrix*

---



메커니즘을 참조하여, 인접 행렬과 차수 행렬의 곱을 계산해 그래프의 연결 정보를 도출한다. 그래프의 연결 정보와 피쳐 정보는 행렬 곱을 통해 합쳐진다. 이 단계는 그래프의 구조를 노드의 특성에 통합시키기 위한 과정으로, 이를 위해 행렬의 차원을 일치시켜야 한다. 첫 번째 GNN레이어의 출력은 그래프의 연결 정보와 함께 곱해져, 두 번째 GNN레이어의 입력으로 사용된다. 마지막으로, 두 번째 GNN레이어의 출력과 원래의 범주형 및 수치형 피쳐 정보가 더해진다. 이 단계는 TabTransformer의 출력과 너무 큰 차이를 발생시키지 않기 위함이다. 최종적으로, MLP 구조의 Readout Layer를 거쳐 모든 노드 정보를 종합하고, 대출 신청자의 채무 불이행 가능성을 이진 분류로 예측한다.

## 4. 데이터 수집 및 전처리

본 연구에서는 미국 최대 P2P 대출 플랫폼인 Lending Club에서 제공하는 데이터를 활용하였다. Lending Club 대출 데이터는 2007년부터 2018년까지 151개의 속성이 있는 2,260,688개의 대출 이력이 포함되어 있다. 데이터 세트에는 대출 기간(36개월 및 60개월)에 따른 대출 상태(전액 지불, 상각, 유예 중, 연체 등)가 포함되어 있다. 본 연구에서 사용하는 대출 이력 데이터 세트의 고객 ID는 익명화 되었으며, 대출 신청자의 인구통계학적 정보 등의 개인 식별 정보는 포함하고 있지 않다. 다음 표 2는 연구에 활용한 주요 변수의 설명이다.

〈표 2〉 변수 설명

변수 유형	변수 명	설명	속성
종속변수	Loan status	현재 대출 상태	범주형
독립변수	Address state	대출 신청서 상 차주의 거주 지역	범주형
	Annual income	차주가 자체 보고한 연간 소득	수치형
	Application type	개인 대출/공동 대출 여부	범주형
	Dept to income	주택담보대출, LC대출을 제외한 총 채무에 대하여 차주의 월 부채 상환액을 자진신고한 월 소득으로 나누어 계산한 비율	수치형
	Earliest credit line	차주의 신용한도가 개설된 최초 월	범주형
	Employment length	고용 기간 (0~10 사이의 수)	수치형
	FICO range high	대출 개시 시 차주의 FICO 상위 경계범위	수치형
	FICO range low	대출 개시 시 차주의 FICO 하위 경계범위	수치형
	Home ownership	주택 소유 상태 (임대, 소유, 저당, 기타)	범주형
	Initial list status	대출의 초기 상장 상태	범주형
	Installment	대출 발생 시 차주가 지불해야 하는 월별 금액	수치형
	Interest rate	대출 이자율	수치형
	Issue date	대출 시행 월	범주형
	Loan amount	차주에게 대출된 총 금액	수치형
	Mortgage account	모기지 계좌 수	수치형
	Open account	차주의 신용보고서에 있는 개방된 신용계좌 수	수치형
	Public records	부정적인 공공 기록의 수	수치형
	Public record bankruptcies	공공 기록의 파산 횟수	수치형
	Purpose	차주의 대출 신청에 따른 대출 구분	범주형
	Revolving Balance	총 리볼빙에 대한 미지급 잔액	수치형
Revolving Utilization	총 리볼빙에 대한 사용자 이용률	수치형	

#### 4.1 데이터 전처리

데이터 세트는 대출 상태가 ‘Fully Paid(전액 지불)’ 또는 ‘Charged Off(상각)’인 대출만 포함하도록 필터링하였다. 데이터 전처리 과정에서 무의미한 속성(예: URL, Member id 등)과 30% 이상의 결측값을 포함하는 속성은 제거하였으며, 특정 변수들에 대하여 아래의 표 3과 같은 데이터 변환이 이루어졌다.

모든 수치형 변수는 최소-최대 정규화를 수행하였으며, transformer 구조의 입력을 위하여, 고유 값을 가진 범주형 변수의 값을 정수로 변환하는 Label Encoding을 수행하였다. 데이터 전처리 수행 후 샘플은 총7개의 범주형 속성과 18개의 수치형 속성으로 구성되어 있다.

### 5. 실험 설계

본 연구에서 제안한 모형은 금융 이력이 부족한 실피파일러의 대출 상환 능력을 평가하기 위해 설계되었다. 기획재정부 정의에 근거하여 실피파일러를 명확히 정의하고, 해당 기준에 따라 실험을 진행하였다. 구체적으로, 최근 2년간 신용카드 실적과 최근 3년간 대출 경험이 없는 사람들을 실피파일러로 분류하였다(기획재정부, 2021).

샘플 추출을 위해, 신용카드 발급 이력과 최근 2년의 신용카드 거래 기록을 통합하여 최근 신용카드 실적을 파악하였다. 또한 대출 발생일 변수를 활용하여 대출 이력을 확인하고, 대출 상태 변수와 연계하여 최근 대출 경험을 파악하였다. 이 두 조건을 모두 만족하는 데이터를 추출하여 총 228,012건의 실피파일러 데이터를 확보하였다.

#### 5.1 모델 구조

본 연구에서 제안한 TeGCN 모델은 Python과 Pytorch 환경에서 구현되었고, 채무 불이행 예측의 정확도를 향상을 위해 하이퍼파라미터를 실험적으로 조정하여 최적값을 선정하였다. 활성화 함수로는 GeLU(Gaussian Error Linear Unit)를 선택하였다. GeLU는 GCN에서 널리 쓰이는 활성화 함수인 RELU에 비해 빠른 수렴 속도와 낮은 오차를 보이는 장점이 있어 사용하였다(Hendrycks & Gimpel, 2016). 손실함수의 최적화를 위해 학습률 0.001의 Adam 알고리즘을 사용하였고, L2 Penalty를 통해 0.001의 가중치 감소를 적용하였다. 또한 모델이 효율적으로 수렴하도록 하기위해 Early Stop 기법을 적용하고, 최대 Epoch를 200으로 설정하였다. 표 4는 TeGCN 모델의 하이퍼파라미터를 제시한다.

〈표 3〉 데이터 전처리 내역

변수 명	데이터 변환
Term	기간의 숫자 값을 추출한 후, 구간화 수행
Employee length	‘10+ years’는 ‘10 years’로, ‘< 1 year’는 ‘0 years’로 설정 후, 정수 변환
Home ownership	‘NONE’, ‘ANY’ 카테고리를 ‘OTHER’라는 단일 카테고리로 병합
Annual income Revolving Balance	데이터 왜곡을 줄이기 위한 로그 변환 적용
Fico score	‘FICO_RANGE_LOW’와 ‘FICO_RANGE_HIGH’의 평균으로 계산
Loan label	‘LOAN_STATUS’로부터 대출 상태 ‘Charged Off’ 여부 파악

〈표 4〉 모델 하이퍼파라미터 설명

하이퍼 파라미터	값
활성화 함수	GeLU
학습률	0.001
최적화 함수	Adam
L2 Penalty	0.001
Early Stopping	20
최대 Epoch	200

## 5.2 베이스라인 모델

모델의 효과를 확인하기 위해, 채무 불이행 예측에 있어 주로 활용되는 Logistic Regression, Random Forest, XGBoost와 같은 머신러닝 기반의 분류 모델과 함께 성능을 비교하였다(Moscato et al., 2021). 더불어, 신용 평가에 자주 적용되는 딥러닝 모델인 다층 퍼셉트론(MLP)과 합성곱 신경망(CNN)과의 성능도 함께 비교하였다(원종관 등 2021). 기준 모델의 자세한 설명과 실험을 위해 설정한 하이퍼파라미터는 다음과 같다.

- Logistic Regression: 이진 분류를 위한 간단하고 효율적인 선형 모델이며, 신용 평가에서 가장 일반적으로 사용되는 전통적인 알고리즘이다(Li et al., 2017).
- Random Forest: 다수의 결정 트리를 훈련하고, 다수결을 통해 레이블을 결정하는 앙상블 학습 방법이다(Breiman, 2001). 본 연구에서 설정한 Random Forest의 트리 수는 100개, 각 트리의 최대 깊이는 제한하지 않았으며, 최대 기능 수는 전체 피쳐 수의 제곱근이다.
- XGBoost: 그래디언트 부스팅 프레임워크를 사용하는 결정 트리 기반 앙상블 기계 학습 알고리즘이다(Chen & Guestrin, 2016). 본 연구

에서 XGBoost의 트리 수는 1200개, 각 트리의 최대 깊이는 6개, 학습률 0.3, 감마 0.098로 설정하였다.

- MLP: 입력 레이어와 출력 레이어 사이에 여러 개의 히든 레이어가 있는 신경망 구조이며, 역전파(back) 알고리즘을 사용하여 손실을 최소화하도록 수행된다. MLP는 학습 데이터의 분포에 관한 가정이 필요하지 않기 때문에, 실제 환경에서 높은 분류 성능을 보인다(Benediktsson et al., 1990). 본 연구에서는 MLP의 레이어 수는 6으로 설정하였으며, 활성화 함수는 ReLU, 학습률은 0.01, 최적화 알고리즘은 SGD(Stochastic Gradient Descent)를 사용하였고, 에포크(epoch)는 200회로 설정하였다.
- CNN: 합성곱 레이어, 풀링 레이어, 완전 연결 레이어로 이루어진 신경망 구조이며, 역전파 알고리즘을 사용하여 학습 가능한 매개변수를 갱신(update)함으로써 네트워크의 학습을 수행한다. CNN은 가중치 공유를 통해 피쳐를 추출하고 학습함으로써, 복잡한 작업에서 높은 성능을 보인다(Yamashita et al., 2018). 본 연구에서 CNN의 활성화 함수는 ReLU, 학습률은 0.001, 최적화 알고리즘은 Adam, 에포크(epoch)는 200회로 설정하였다.
- MLP+GCN: 입력 레이어와 그래프 합성곱 레이어로 구성된 신경망 구조이며, 역전파 알고리즘을 사용하여 손실을 최소화하도록 학습한다. 본 연구에서는 Tab Transformer를 활용한 범주형 변수 임베딩의 효과를 측정하기 위해 TeGCN에서 Tab Transformer를 제외하고, MLP 레이어를 추가하였다. 활성화 함수는 ReLU, 학습률은 0.001, 최적화 알고리즘은 Adam, 에포크(epoch)는 200회로 설정하였다.

### 5.3 성능 평가 지표

기준 모형과 제안 모형의 성능을 비교하기 위해 AUC(Area Under the Curve)와 K-S(Kolmogorov-Smirnov) 통계량을 확인하였다. AUC는 ROC (Receiver Operator Characteristic) 곡선 아래의 영역으로 표현된다. 이진 분류 문제에서 ROC 곡선은 X축이 1-특이도, Y축이 민감도인 그래프이며, AUC 값은 1에 가까울수록 분류 모형의 성능이 좋다는 것을 의미한다. K-S 통계량은 두 집단의 누적 분포의 차이를 나타내는 지표로, 대상이 ‘채무 이행’과 ‘채무 불이행’으로 구분된 신용평가 모형의 성능 평가 시 주로 활용된다(홍중선, 김지훈, 2009).

## 6. 실험 결과

본 장에서는 일반 대출 신청자 및 실패율 데이터 세트에서 베이스라인 모형과 TeGCN 모형 간 성능 비교를 수행했다. 이때, 모형의 일반화 가능성을 검증하기 위해 K-Fold 교차검증을 수행하였다. 교차검증은 총 5회 실시하였으며, 각 Fold의 평균을 통해 도출한 실험 결과는 표 5와 같다. TeGCN 모형은 일반 대출 신청자 데이터에 대하여 5개의 머신러닝 및 딥러닝 기반 베이

스라인 모형보다 성능이 뛰어난 것으로 나타났다. AUC 값과, K-S 통계량 두 평가 지표에서 모두 TeGCN 모형이 우수한 것으로 나타났으며, 이는 그래프 정보를 활용한 신용 평가 모형이 채무 불이행 예측이 유의미한 것을 나타낸다. 또한, 금융 이력이 부족한 실패율 데이터 세트에서 역시 베이스라인 모형 대비 높은 성능을 보였다. 특히, 일부 베이스라인 모형들이 실패율 예측 시 성능이 떨어진데 반해, 본 모형은 비교적 안정적인 성능을 유지했다. 전체 데이터 세트와 마찬가지로 실패율 데이터에서도 기준 모형보다 TeGCN 모형의 성능이 뛰어난 것으로 나타났다. 마지막으로, TeGCN은 Tab Transformer 대신 MLP 레이어를 추가한 GCN 모형 대비 뛰어난 성능을 기록했다. 이는 Tab Transformer를 활용한 범주형 변수 임베딩이 채무 불이행 예측 성능 향상에 유의미한 도움이 된다는 것을 의미한다.

추가적으로, 타겟 클래스 균형에 대한 모형의 영향과 성능을 확인하기 위하여 소수 클래스와 다수 클래스의 비율을 1:1로 맞춰 모형에 대한 Sub-Sample 테스트를 아래 표 6과 같이 진행하였다. 그 결과, 클래스 균형의 상황에서도 모형은 기존 모형 대비 준수한 성능을 보였다.

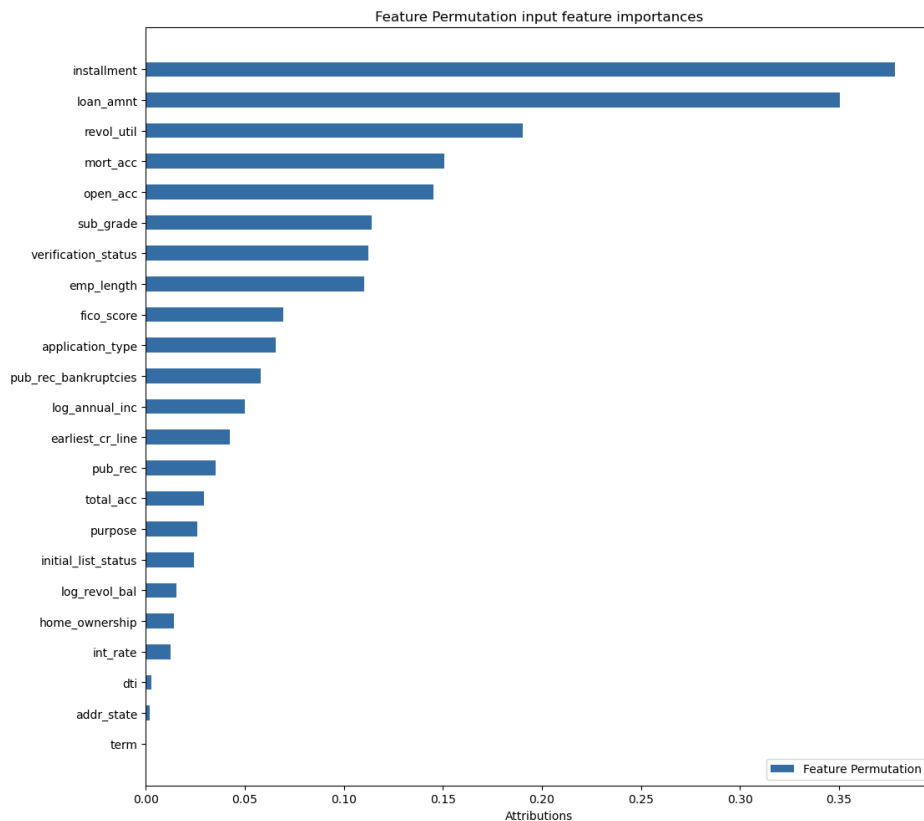
〈표 5〉 모형 실험결과

Model	전체 데이터		실패율 데이터	
	AUC	KS Score	AUC	KS Score
Logistic Regression	0.69928	0.28425	0.67597	0.26192
Random Forest	0.70607	0.29871	0.70903	0.30673
XGBoost	0.72107	0.32093	0.71563	0.31609
MLP	0.71180	0.30566	0.70798	0.27965
CNN	0.71245	0.34288	0.71637	0.35279
MLP+GCN	0.70110	0.29668	0.69108	0.27808
<b>TeGCN (Our model)</b>	<b>0.77028</b>	<b>0.45842</b>	<b>0.77430</b>	<b>0.46701</b>

〈표 6〉 Sub-Sample 실험결과

Model	신파일러 데이터 (Down Sampling)	
	AUC	KS Score
Logistic Regression	0.67051	0.25596
Random Forest	0.69998	0.29774
XGBoost	0.69766	0.28580
MLP	0.65130	0.27254
CNN	0.70094	0.31226
MLP+GCN	0.70109	0.30394
<b>TeGCN (Our model)</b>	<b>0.75697</b>	<b>0.43239</b>

본 연구는 모델 학습에 활용된 23개의 핵심 변수들의 중요도를 분석하였다. 중요도 평가를 위하여 Feature Permutation 방식을 적용하였다. 이 방법은 특정 변수의 값을 임의로 섞어, 기존 출력과 변경된 출력 사이의 차이로 해당 변수의 중요도를 측정한다. 만약 해당 변수 값을 변경해도 모델 출력에 크게 영향을 미치지 않는다면, 그 변수는 상대적으로 중요하지 않다는 해석이 가능하다. 그러나 큰 차이가 발생한다면, 해당 변수는 모델에서 핵심적인 역할을 수행한다고 볼 수 있다. Feature Permutation 방식을 통해 확인한 변수 중요도는 그림 3과 같다. ‘분할 납부금(installment)’과



〈그림 3〉 변수 중요도 그래프

‘대출 금액(loan\_amount)’ 등의 주요 대출 관련 변수는 채무 불이행 예측에 결정적인 영향을 주는 것으로 나타났다. 그 외에도 ‘리볼빙 서비스 이용률(revlo util)’과 ‘모기지 계좌 수(mort\_acc)’ 그리고 ‘신용계좌 수(open\_acc)’와 같은 변수들 또한 중요한 역할을 하는 것으로 평가되었다.

## 7. 결론

본 연구는 TabTransformer와 GCN을 결합한 TeGCN 채무 불이행 예측모델을 제안하였다. TabTransformer 구조를 활용하여 신용평가 데이터에 존재하는 다양한 범주형 변수의 영향력을 적절히 반영하였으며, GCN 구조를 통해 씬파일러의 데이터 부족 문제를 완화하였다. 실험 결과, 본 모델은 일반 대출 신청자와 씬파일러 데이터셋 모두에 대하여 채무 불이행에 활용되는 타 모델 대비 우수한 성능을 보였다. 특히, 일부 베이스라인 모델들이 씬파일러의 채무 불이행 예측 시 성능이 떨어진데 반해, TeGCN 모델은 안정적인 성능을 유지했다.

채무 불이행 예측에 활용되는 기존 모델들에 비해 높은 성능을 달성했음에도 불구하고, 본 연구는 몇 가지 한계가 존재한다. 첫째, 학습에 활용한 데이터의 일반화 가능성을 추가적으로 검증할 필요가 있다. TeGCN 모델의 학습에는 미국의 P2P 대출 플랫폼인 Lending Club의 대출이력 데이터가 활용되었다. Lending Club 데이터의 경우 미국의 대출 플랫폼이며, P2P라는 독특한 특성을 갖기에 국내의 대규모 금융사에서 본 모델을 활용 가능한지 추가적인 검증이 필요하다. 둘째, 본 연구에서는 차주간의 네트워크 관계를 차주의 실제 값의 유사도를 바탕으로 정의하였다. 해당 방식은 계산 복잡도를 낮춘다는 장점을 가지고 있지만, 차

주간의 관계를 정의하기 위해 다양한 방식을 추가적으로 고려할 수 있다. 예를 들어, 데이터 포인트의 거리를 바탕으로 한 KNN-Graph, K-means와 같은 방식과 함께 Transformer를 활용한 Relation Encoder 방식을 활용할 수 있다. 따라서, 새로운 네트워크 관계 정의 방식에 대한 모델 성능 검증을 추가적으로 진행할 필요가 있다. 마지막으로, 현재 실무에서는 씬파일러 신용평가를 위해 이커머스의 주문내역 등 비금융정보를 수집 및 활용하고 있다(김종윤 등, 2019). 이러한 비금융정보를 변수로 추가한다면 모델의 성능을 더 높일 수 있을 것이다.

본 연구의 학술적, 실무적 시사점은 다음과 같다. 먼저, TeGCN 모델은 범주형 변수가 많은 신용정보의 특성을 고려하여 Transformer 기반의 임베딩 방법론을 적용하였으며, 이를 통해 채무 불이행 예측에 활발히 사용되는 베이스라인 모델 대비 높은 성능을 달성하였다. 범주형 변수를 원핫 인코딩으로 표현하던 기존의 연구들과 달리, 새로운 범주형 변수 임베딩 방법론을 발굴하고 적용하였으며, 유의미한 성능 향상을 거두었다는 점에서 학술적 의의를 갖는다. 또한, 금융 이력 정보가 부족한 씬파일러의 채무 불이행을 예측하기 위해 그래프 기반 모델을 적용하였다는 점에서 신용평가 모델에 대한 기존 연구범위를 확장하였다. 실무적 의의로는 씬파일러의 금융 소외문제를 완화하고 금융사의 추가적인 수익 기회를 창출했다는 점이 있다. 본 연구에서 제안한 TeGCN 모델은 씬파일러의 채무 불이행을 정확히 예측함으로써 금융사들이 씬파일러 대상 금융 서비스를 확대하는데 도움을 줄 수 있을 것이다.

## 참고문헌(References)

### [국내 문헌]

- 권영탁. (2021, 08.17). [권영탁의 핀테크 이야기] 대안신용평가로 핀파일러들 구제해야. 한국금융신문. [https://m.fntimes.com/html/view.php?ud=202108140937274397dd55077bc2\\_18](https://m.fntimes.com/html/view.php?ud=202108140937274397dd55077bc2_18)
- 김명중. (2012). 회사채 신용등급 예측을 위한 SVM 앙상블학습. *지능정보연구*, 18(2), 29-45.
- 김성진, & 안현철. (2016). 기업신용등급 예측을 위한 랜덤 포레스트의 응용. *산업혁신연구*, 32(1), 187-211.
- 김연정. (2021, 09.27). ‘금융이력부족자’ 대부분 신용점수 700점대...은행 대출에 불리. 연합뉴스. <https://www.yna.co.kr/view/AKR20210926043100002>
- 김종윤, 장원중, & 김광용. (2019). 온라인 상거래 데이터를 반영한 개인신용평가모형 (커머스 스코어) 개발. *정보기술아키텍처 연구*, 16(1), 45-55.
- 김정산. (2023, 05.09). 금융소비자 양극화...“저신용자 더 빌리고 고신용자 더 모았다”. 메트로신문. <https://www.metroseoul.co.kr/article/20230509500001>
- 김하영, 허정윤, & 권호창. (2022). 인공지능 기반 금융서비스의 공정성 확보를 위한 체크리스트 제안: 인공지능 기반 개인신용평가를 중심으로. *지능정보연구*, 28(3), 259-278.
- 엄하늘, 김재성, & 최상욱. (2020). 머신러닝 기반 기업부도위험 예측모델 검증 및 정책적 제안: 스테킹 앙상블 모델을 통한 개선을 중심으로. *지능정보연구*, 26(2), 105-129.
- 오운실. (2020). 신용등급 재평가를 통한 대출 상황 여부 예측. *한국정보과학회 학술발표논문집*, 580-582.
- 유성준, & 박나리. (2020). CART 기법을 이용한

개인신용정보 재현자료 생성 기법. *통계연구*, 25(1), 1-30.

- 이근희, 유영범, & 하승인. (2017). 개인신용평가 모형을 위한 딥러닝 활용에 대한 연구. *대한산업공학회 춘계공동학술대회 논문집*, 4042-4047.
- 이우형. (2018). 딥러닝 학습의 판별 성능 증대를 위한 부스팅 활용에 대한 연구 : 신용평가 모형을 중심으로. 서강대학교.
- 김유진. (2022, 01.03). [신년기획] 가계대출 총량 관리 시작... 은행, 35조 증금리 대출시장서 기회 노린다. 이투데이, <https://www.EToday.co.kr/news/view/2092457>
- 김나경 & 김예지. (2023, 03.15). [단독] ‘無이력의 악순환’ 핀파일러 99.9%가 신용 800점 이하... “거래이력이 없지, 빚 안 갚는다 했다”. 파이낸셜뉴스, <https://www.fnnews.com/news/202303151401284618>
- 홍종선, & 김지훈. (2009). 신용평가모형에서 두 분포함수의 동일성 검정을 위한 비모수적인 검정방법. *한국데이터정보과학회지*, 20(2), 261-272.

### [국외 문헌]

- Abrahams, C. R., & Zhang, M. (2008). Fair lending compliance: Intelligence and implications for credit risk management (Vol. 13). John Wiley & Sons.
- Agosto, A., Giudici, P., & Leach, T. (2019). Spatial regression models to improve P2P credit risk management. *Frontiers in artificial intelligence*, 2, 6.
- Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, 49(3), 312-329.
- Benediktsson, J. A., Swain, P. H., & Ersoy, O. K.

- (1990). Neural network approaches versus statistical methods in classification of multisource remote sensing data. Vancouver, Canada, July 10-14, 1989) IEEE Transactions on Geoscience and Remote Sensing.
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- Calabrese, R., Elkin, J. A., & Giudici, P. S. (2017). Measuring bank contagion in Europe using binary spatial regression models. *Journal of the Operational Research Society*, 68, 1503-1511.
- Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107 (8-10), 1477-1494.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.
- Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access*, 9, 114381-114391.
- Duarte, F., Martins, B., Pinto, C. S., & Silva, M. J. (2018). Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *Journal of biomedical informatics*, 80, 64-77.
- Fu, X., Ouyang, T., Chen, J., & Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing & Management*, 57(4), 102236.
- Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.
- He, H., & Fan, Y. (2021). A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. *Expert Systems with Applications*, 176, 114899.
- Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.
- Huang, X., Khetan, A., Cvitkovic, M., & Karmin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678.
- Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
- Lagasio, V., Pampurini, F., Pezzola, A., & Quaranta, A. G. (2022). Assessing bank default determinants via machine learning. *Information Sciences*, 618, 87-97.
- Lee, J. W., & Sohn, S. Y. (2021). Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. *PloS one*, 16(12), e0261737.
- Lee, J. W., Lee, W. K., & Sohn, S. Y. (2021). Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. *Expert Systems with Applications*, 168, 114411.
- Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74, 105-114.
- Li, Z., Wang, X., Yao, L., Chen, Y., Xu, G., & Lim, E. P. (2022, November). Graph Neural Network with Self-attention and Multi-task



- Learning for Credit Default Risk Prediction. In *Web Information Systems Engineering - WISE 2022: 23rd International Conference, Biarritz, France, November 1 - 3, 2022, Proceedings* (pp. 616-629). Cham: Springer International Publishing.
- Liu, X., Li, Y., Jiang, C., Wang, Z., Zhao, F., & Wang, J. (2022, May). Attentive feature fusion for credit default prediction. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 816-821). IEEE.
- Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65, 465-470.
- Moscato, V., Picariello, A., & Sperli, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, 113986.
- Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023). On the combination of graph data for assessing thin-file borrowers' creditworthiness. *Expert Systems with Applications*, 213, 118809.
- Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.
- Shumovskaia, V., Fedyanin, K., Sukharev, I., Berestnev, D., & Panov, M. (2021). Linking bank clients using graph neural networks powered by rich transactional data. *International Journal of Data Science and Analytics*, 12, 135-145.
- Vong, W. K., Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2019). Do additional features help or hurt category learning? The curse of dimensionality in human learners. *Cognitive science*, 43(3), e12724.
- Woo, H., & Sohn, S. Y. (2022). A credit scoring model based on the Myers - Briggs type indicator in online peer-to-peer lending. *Financial Innovation*, 8(1), 1-19.
- Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9, 611-629.
- Zhang, L., Wang, J., & Liu, Z. (2023). What should lenders be more concerned about? Developing a profit-driven loan default prediction model. *Expert Systems with Applications*, 213, 118938.
- Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1), 1-23.
- Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 10(6), 1519-1525.
- Zhou, X., Zhang, W., & Jiang, Y. (2020). Personal credit default prediction model based on convolution neural network. *Mathematical Problems in Engineering*, 2020, 1-10.

#### [URL]

- 기획재정부. (2021.02.17). *신파일러 정의*. 09.05, 2023, from <https://www.moef.go.kr/sisa/dictionary/detail?idx=3216>.
- 네이버 테크핀 리포트. (2021, 09.04). 09.05, 2023, from [https://www.navercorp.com/navercorp/research/2022/20220217202609\\_2.pdf](https://www.navercorp.com/navercorp/research/2022/20220217202609_2.pdf).
- 통계청. (2023, 03.22). *성별 연령대별 소득*. 09.05, 2023, from [https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT\\_1EP\\_2010&conn\\_path=12](https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1EP_2010&conn_path=12)

## Abstract

# TeGCN:Transformer-embedded Graph Neural Network for Thin-filer default prediction

Seongsu Kim\* · Junho Bae\* · Juhyeon Lee\* · Heejoo Jung\*\* · Hee-Woong Kim\*\*\*

As the number of thin filers in Korea surpasses 12 million, there is a growing interest in enhancing the accuracy of assessing their credit default risk to generate additional revenue. Specifically, researchers are actively pursuing the development of default prediction models using machine learning and deep learning algorithms, in contrast to traditional statistical default prediction methods, which struggle to capture nonlinearity. Among these efforts, Graph Neural Network (GNN) architecture is noteworthy for predicting default in situations with limited data on thin filers. This is due to their ability to incorporate network information between borrowers alongside conventional credit-related data. However, prior research employing graph neural networks has faced limitations in effectively handling diverse categorical variables present in credit information. In this study, we introduce the Transformer embedded Graph Convolutional Network (TeGCN), which aims to address these limitations and enable effective default prediction for thin filers. TeGCN combines the TabTransformer, capable of extracting contextual information from categorical variables, with the Graph Convolutional Network, which captures network information between borrowers. Our TeGCN model surpasses the baseline model's performance across both the general borrower dataset and the thin filer dataset. Specially, our model performs outstanding results in thin filer default prediction. This study achieves high default prediction accuracy by a model structure tailored to characteristics of credit information containing numerous categorical variables, especially in the context of thin filers with limited data. Our study can contribute to resolving the financial exclusion issues faced by thin filers and facilitate additional revenue within the financial industry.

**Key Words** : Thin Filer, Default Prediction, Graph Convolutional Network, Categorical Feature Embedding, TeGCN

Received : September 6, 2023 Revised : September 22, 2023 Accepted : September 24, 2023

Corresponding Author : Hee-Woong Kim

---

\* Graduate School of Information, Yonsei University

\*\* Hana TI, Cloud Center MSP Cell

\*\*\* Corresponding author: Hee-Woong Kim

Graduate School of Information, Yonsei University

50, Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

Tel: +82-2 2123-4195, Fax: +82-2 2123-4195, E-mail: kimhw@yonsei.ac.kr

## 저자 소개



**김성수**

연세대학교 정보대학원에서 정보시스템 석사과정에 재학중이다. 현재 디지털서비스연구실에서 인과추론 관련 연구들을 진행 중이다. 관심 분야는 Information Retrieval, Causal inference, Deep Learning 등이다.



**배준호**

연세대학교 정보대학원 정보시스템 석사과정이다. 데이터 분석가로 재직한 경력이 있다. 관심 연구 분야는 Natural Language Processing, Anomaly Detection, Deep Learning Application 등이다.



**이주현**

연세대학교 정보대학원에서 정보시스템 석사과정에 재학 중이다. 현대백화점에서 마케팅으로 재직 한 경력이 있다. 관심 분야는 Natural Language Processing in Deep Learning, Time Series Analysis 등이다.



**정희주**

하나금융티아이에서 클라우드/가상화 기반 인프라 구축 및 서비스 개발을 담당하고 있다. 주요 관심 분야는 OpenStack 기반 클러스터 구축 및 운영이며, Anomaly Detection, 네트워크 인프라 설계 등이다.



**김희웅**

National University of Singapore 정보시스템학과에서 근무한 후, 현재 연세대학교 정보대학원 교수로 재직 중이다. 주요 연구분야는 디지털 비즈니스, 정보시스템 관리 및 활용 등이다. 관련 연구들은 MIS Quarterly, Information Systems Research, Journal of Management Information Systems 등에 60여 편의 논문이 게재되었다. MIS Quarterly, JAIS, IEEE TEM의 편집위원으로 활동했고, KrAIS회장을 역임했다.