순차적 데이터 결합 방법을 통한 이종 산업 간 추천 시스템 문제 해결

Solving Problems in Cross-Industry Recommendation System Through Sequential Data Combination Method

오현진(한국교통대학교 소프트웨어학전공), **정수민**(경북대학교 통계학과), **정은혜**(석사과정, 성신여자대학교 심리학과), **조수현**(지도교수, 명지대학교 AI정보과학전공)

Introduction

- 최근 다양한 분야간 데이터 결합 및 활용의 활성화를 통해 추천 시스템 개발, 정책 수립 등을 효과적으로 수행하고자 하는 움직임이 정부 주도 하에 활발히 진행되고 있다.
- 기존의 데이터 결합 방식은 정보 제공 주체의 민감한 정보를 직접 공유하기 때문에, 개인정보 노출 등의 위험 부담을 가진다.
- 이러한 문제를 해결하기 위해 개인정보 보호를 강화할 수 있는 데이터 결합 방법에 대한 관심이 높아지고 있다.

Contributions

- 본 연구자들은 성별, 나이 등 최소한의 고유정보가 포함된 익명 데이터를 활용한 순차적 데이터 결합 방법과, 통계적 데이터 결합 방법을 제안한다.
- 결합 대상 데이터로 보험료 연체 데이터와 자동차 보험 관심도 데이터를 사용하였다.
- 모델의 학습 과정에서 데이터 보안 강화를 위해 분할 학습(Split Learning)과 연합 학습(Federated Learning) 방식을 사용하였다.
- 대표적인 XAI(설명 가능한 인공지능) 방법론인 SHAP의 Shapley Value를 활용해 각 feature의 기여도를 계산, 분류 결과에 대한 설명을 제공하였다.

Methods

1. 순차적 데이터 결합

- 성별, 나이와 같은 고유 정보를 기준으로 첫 번째 데이터셋(신용카드 연체 예측 데이터)을 정렬한 후, 두 번째 데이터셋(건강 보험 관심도 데이터)에서 동일한 고유 정보를 가진 데이터를 순차적으로 반복하여 결합하였다.

그림1. 순차적 데이터 결합 방식

Client A			
A	남	21세	100
B	남	21세	200
C	남	21세	300
D	남	21세	400

Client B			
가	남	21세	50
나	남	21세	500

➡

Virtual Data				
A	남	21세	100	50
B	남	21세	200	500
C	남	21세	300	50
D	남	21세	400	500

2. 통계적 데이터 결합

- 하나의 데이터셋 내에서 성별, 나이와 같은 고유 정보를 기준으로 그룹을 생성한 후, 각 그룹별로 변수에 대한 통계량(예: 평균, 최빈값, 중앙값 등)을 집계하였다.
- 이후 결합하고자 하는 데이터셋에서 동일한 최소 고유 정보를 가진 그룹에게 해당 통계량을 새로운 변수로 추가하였다.

그림2. 통계적 데이터 결합 방식

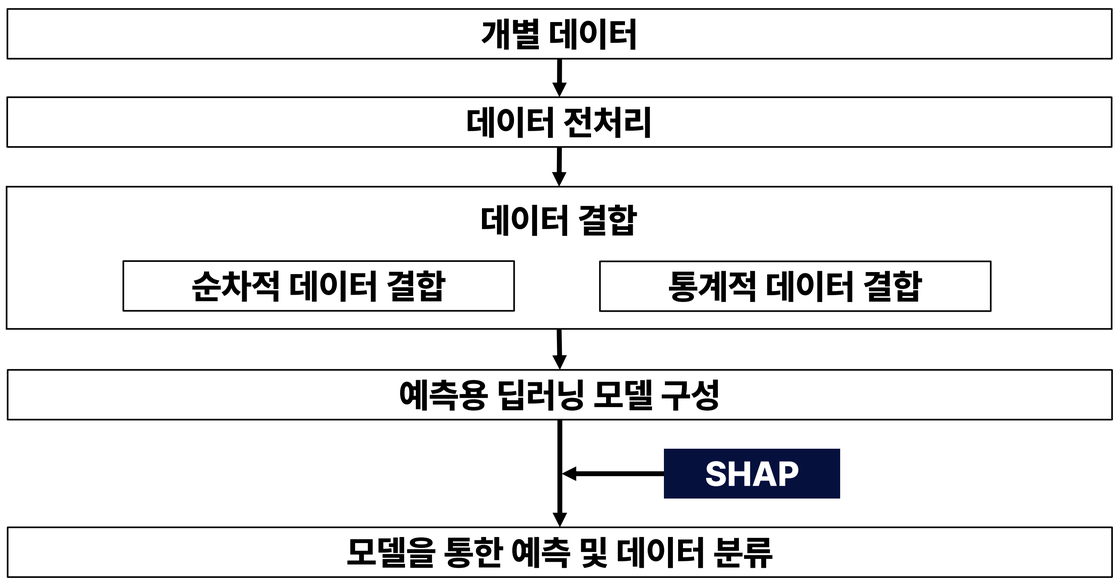
Client A			
A	남	21세	100
B	남	21세	200
C	여	21세	300
D	여	24세	400

Client B			평균	최빈	중앙
1	남	21세	15	17	14
2	여	21세	20	15	18
3	여	24세	7	7	7

➡

Virtual Data				평균	최빈	중앙	...
A	남	21세	100	15	17	14	...
B	남	21세	200	15	17	14	...
C	여	21세	300	20	15	18	...
D	여	24세	400	7	7	7	...

3. Modeling



Results

표1. 데이터별 학습 성능 평가 지표

	Accuracy	Precison	Recall	F1 Score
신용카드 연체 예측 데이터	0.49	0.16	0.33	0.22
신용카드 연체 예측 데이터 (순차적 데이터 결합)	0.50	0.44	0.39	0.35
신용카드 연체 예측 데이터 (통계적 데이터 결합)	0.51	0.32	0.38	0.31
건강보험 관심도 예측 데이터	0.68	0.40	0.00	0.00
건강보험 관심도 예측 데이터 (순차적 데이터 결합)	0.76	0.58	0.76	0.66
건강보험 관심도 예측 데이터 (통계적 데이터 결합)	0.76	0.60	0.71	0.65

- 동일한 딥러닝 모델에 기존의 단일 데이터와 순차적, 통계적 데이터 결합을 통해 결합한 데이터를 학습시킨 후 비교하였으며, 그 결과는 표 1과 같다.
- Accuracy를 비교해 보았을 때는 성능의 차이가 크게 없는 것처럼 보이나, 나머지 지표들을 함께 비교해보면 유의미한 차이가 발생한 것을 확인할 수 있었다.
- 더불어 단일 데이터로 학습한 모델은 단일 y값에 편향된 예측을 하는 반면, 결합한 데이터로 학습한 모델은 모든 y값을 골고루 예측하는 것을 확인할 수 있었다.

그림3. 순차적으로 결합된 신용카드 연체 데이터에 대한 Shapley Values

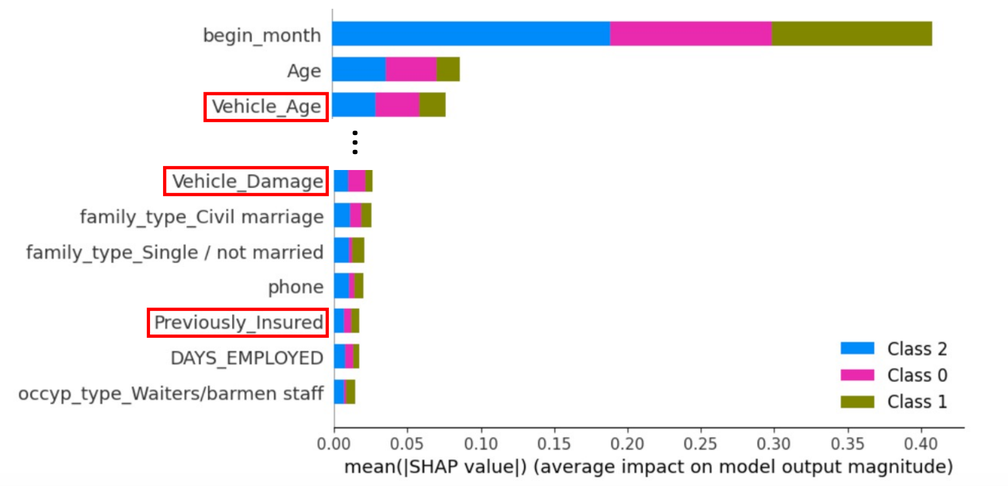
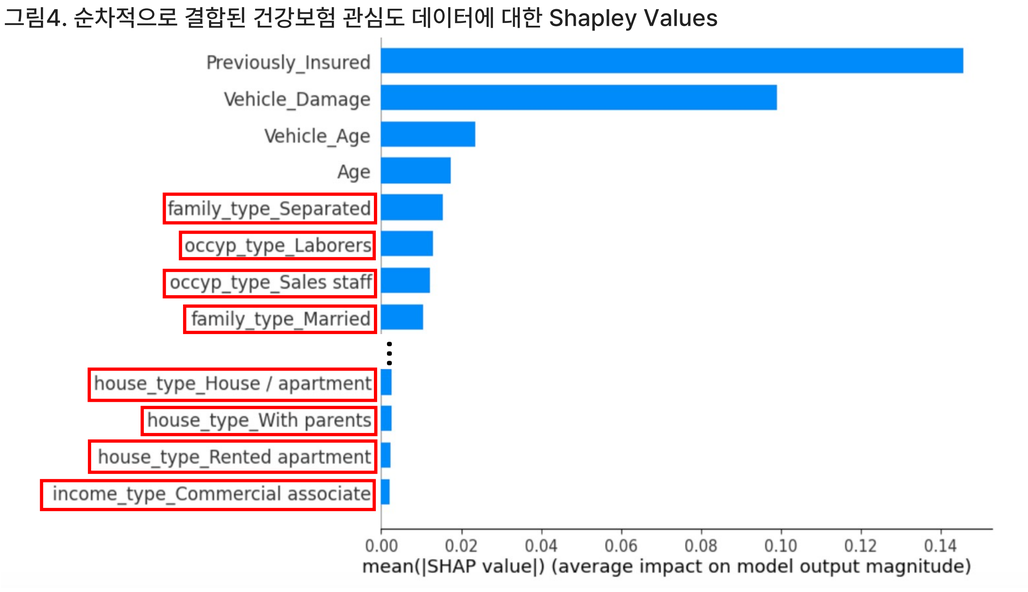


그림4. 순차적으로 결합된 건강보험 관심도 데이터에 대한 Shapley Values



- 두 그래프에서 공통적으로 기존 데이터의 features 뿐 아니라, 결합한 데이터의 features 또한 상당히 중요한 요소로 나타난다는 것을 확인할 수 있다.
- 이는 데이터 결합 이후에도 학습에 기존 데이터만이 영향을 미치는 것이 아닌, 결합한 데이터 또한 중요한 요소로 사용되어 학습에 영향을 행사하는 것으로 해석할 수 있다.

Conclusions

- 순차적 데이터 결합을 통해 생성된 데이터로 예측을 수행하는 경우, 단일 데이터를 통해 예측을 수행하는 경우에 비해 정확도 및 y값 편향 문제에 있어 향상된 성능을 보였다.
- 순차적 데이터 결합은 성별, 나이와 같은 최소한의 개인정보만을 이용하여 데이터를 결합하므로, 데이터 결합 과정에서 우려되는 개인정보 노출 문제에 대한 해결 방안으로써 다양한 산업군의 활발한 데이터 결합에 기여할 수 있다.