YONSEI UNIVERSITY

Using Federated learning instead of imputing missing values

Team NOMAD,

김종현(2022311815), 양희민(2022311911),
이주현(2022311873), 최영조(2022311917)

Severance

Research question

- 데이터에 결측치가 있는 경우 학습이 어려워지고, 결측치를 임의로 채우려 하면 오히려 데이터에 대한 신뢰도가 떨어질 수 있다는 문제가 있음

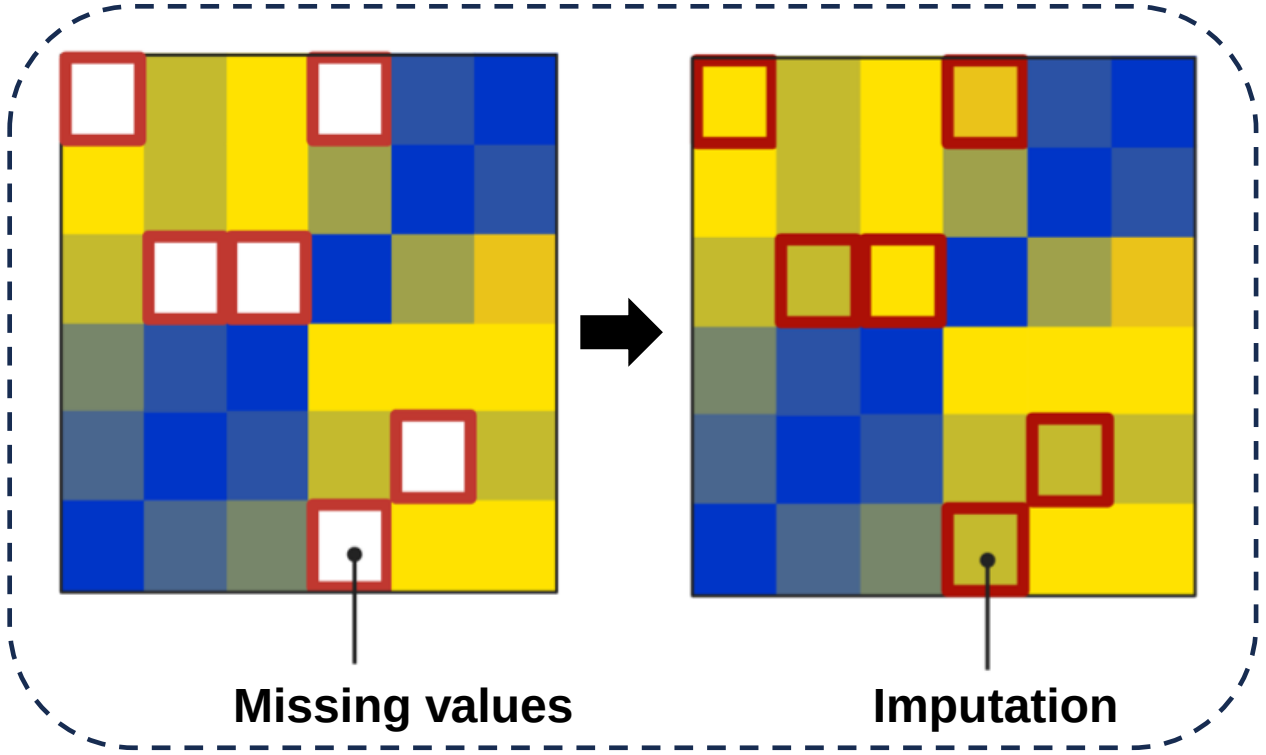
- 결측치를 채우거나 결측치에 대한 전처리 작업을 진행하지 않아도 성능이 우수하거나 비슷한 결과를 보일 수 있는지를 확인해보고자 함

Introduction

- **Missing values (결측치)**

-> 결측치는 dataset에서 누락된 값 또는 측정되지 않는 값을 의미

< Preprocess of missing values >



Missing values의 문제점

- Machine learning 알고리즘은 missing values를 처리하지 못하는 경우가 많음
- Missing values 처리 과정에서 정보 loss 나 bias 가 발생할 수 있음

Introduction

- **Missing values 처리**

1) Listwise Deletion (행 제거)
2) Column Deletion (열 제거)
3) Mean Imputation (평균값 대체)
4) Median Imputation (중앙값 대체)
5) Mode Imputation (최빈값 대체)
6) K-Nearest Neighbors Imputation (K-최근접 이웃)

Related work

● Federated Learning (연합 학습)

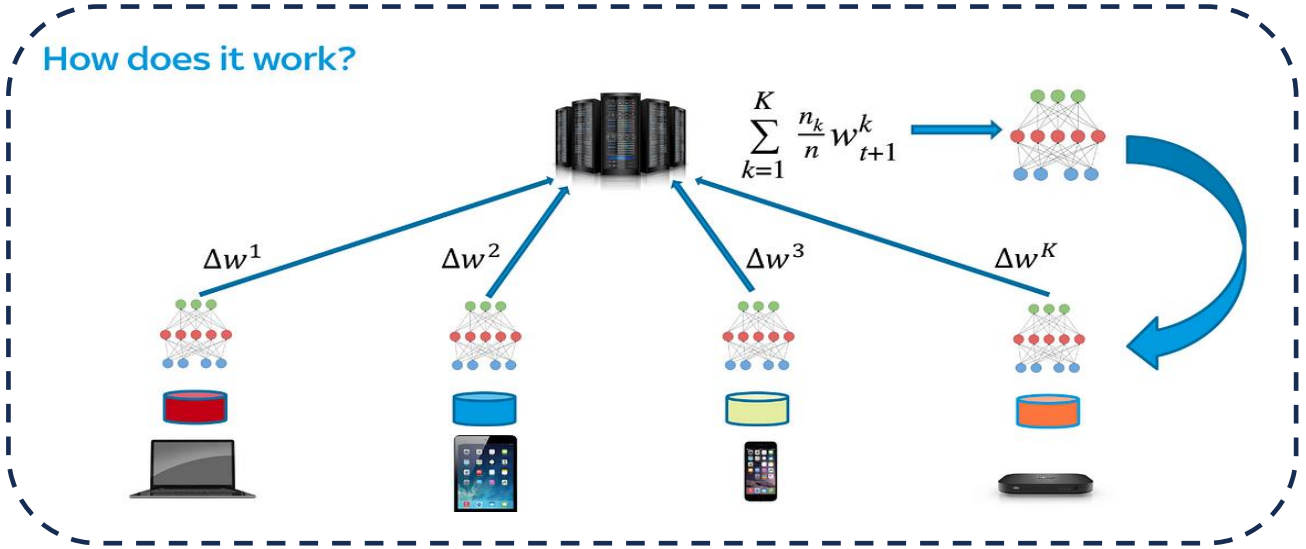
- 데이터를 한 곳에 모을 수 없는 상황에서 모델을 학습시키는 방법

- 여러 local (Client)에서 각 모델을 학습 후, 학습된 parameter server에서 취합

  -> Aggregation 후 각 local로 분배

  -> local 에서 다시 학습 후 server에서 aggregation 반복

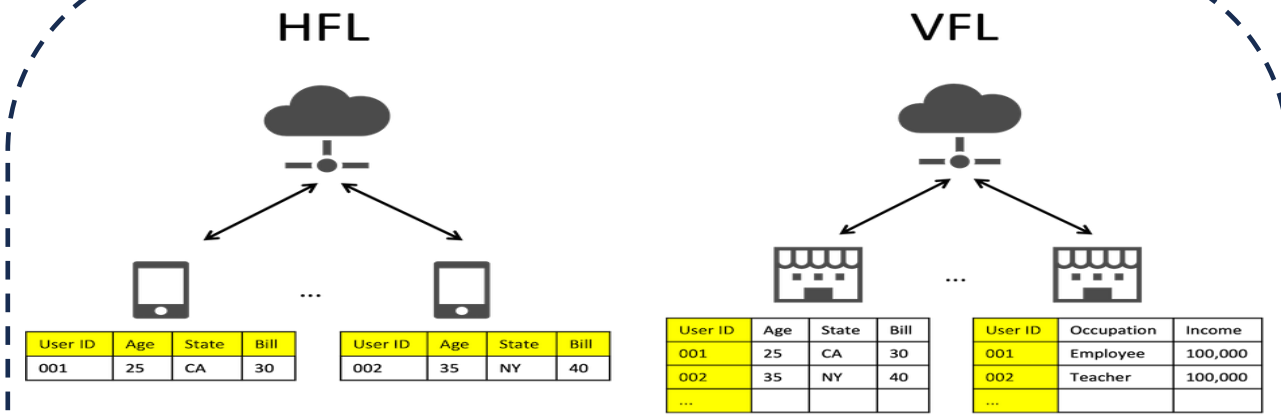
< Federated learning workflow >



Related work

< Federated learning approach type>

Horizontal Federated Learning | **Vertical Federated Learning**



	HFL	VFL
Data Difference	Sample Space	Feature Space
Scenario	Cross-device / Cross-silo	Cross-silo
Exchanged	Model Parameter	Intermediate results
Kept in local	Local Data	Local Data and Model
Each Party obtains	A Shared Global Model	A Local Model

Objective

- Missing value가 있는 데이터에 대해 Federated learning을 이용한 학습
- Imputation의 불확실성 제거

Method

● Dataset

- **Bank Account Fraud Dataset Suite (NeurIPS 2022)**
  - ➢ 은행계좌 사기 데이터 세트
  - ➢ 총 6개의 서로 다른 은행 계좌 사기 tabular 데이터
  - ➢ Data size: 1,000,000 rows x 31 columns
  - ➢ 명목형 변수 : label encoding
  - ➢ 연속형 변수 : Z-score normalization
  - ➢ Train test split

< 명목형 변수 >

```
payment_type                                {AC, AD, AB, AA, AE}
employment_status                  {CE, CF, CD, CA, CB, CG, CC}
housing_status                     {BF, BG, BE, BB, BD, BA, BC}
source                                        {INTERNET, TELEAPP}
device_os              {macintosh, windows, linux, x11, other}
dtype: object
```

< 종속 변수 분포 >

종속변수 비율	원본	Under-sampling	Train set
0	988,971	11,029	7,720
1	11,029	11,029	7,720

https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?resource=download

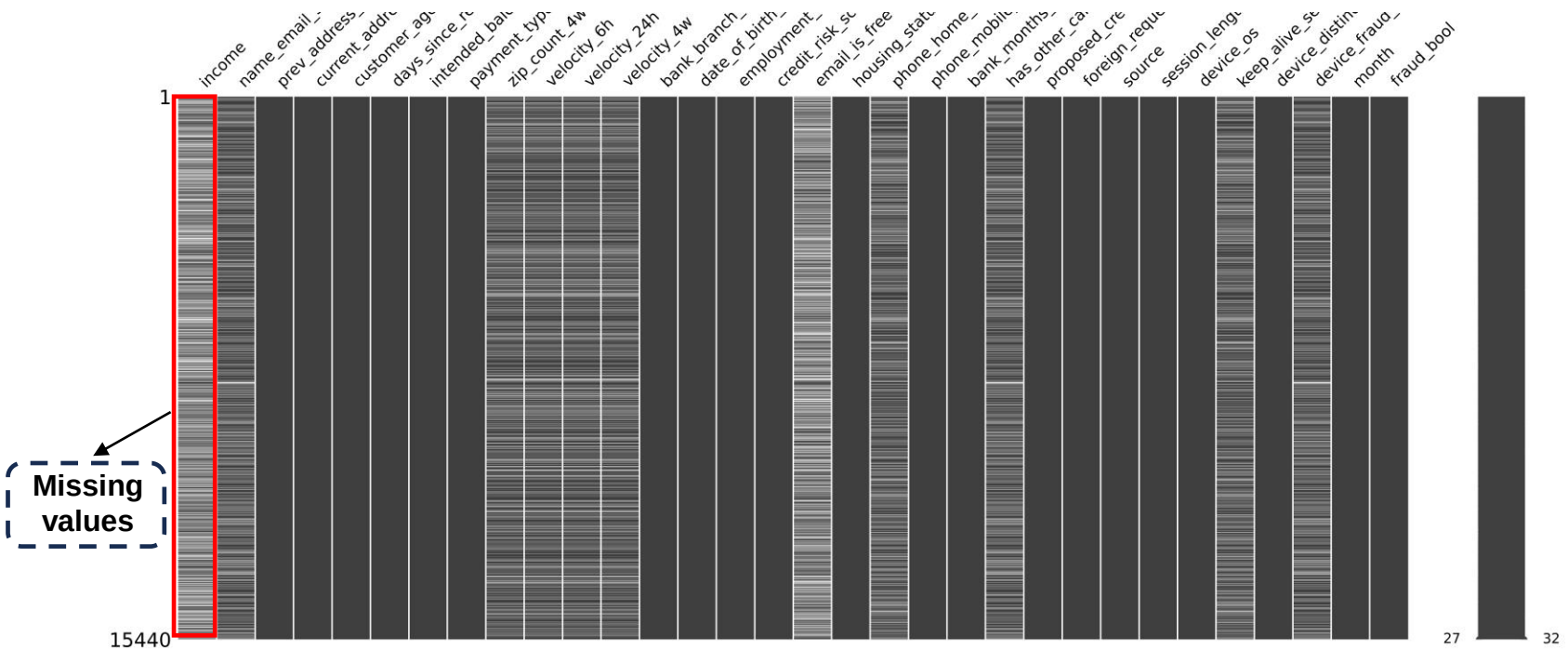
Method - Dataset 구성

Data	Data feature	Data	Data feature
Income	numeric	Employment status	categorical
Name email similarity	numeric	Credit risk score	numeric
Previous address months count	numeric	Email is free	binary
Current address months count	numeric	Housing status	categorical
Customer age	numeric	Phone home valid	binary
Days since request	numeric	Phone mobile valid	binary
Intended balcon amount	numeric	Bank months count	numeric
Payment type	categorical	Has other cards	binary
Zip count 4weeks	numeric	Proposed credit limit	numeric
Velocity 6hours	numeric	Foreign request	binary
Velocity 24hrous	numeric	source	categorical
Velocity 4weeks	numeric	Session length in minutes	numeric
Bank branch count 8weeks	numeric	Device os	categorical
Data of birth distinct emails 4week	numeric	Keep alive session	binary
Device fraud count	numeric	Fraud bool	
month	numeric		

Method

● Customize the dataset (making missing values)

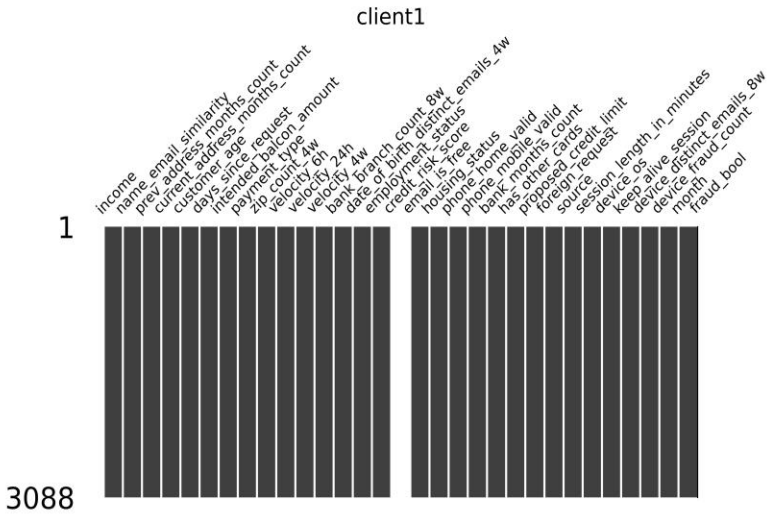
- 해당 데이터셋에 대해 의도적으로 결측치를 생성함
- 생성된 결측치를 고려하여 클라이언트를 분활하고, 이에 대한 연합 학습을 적용함

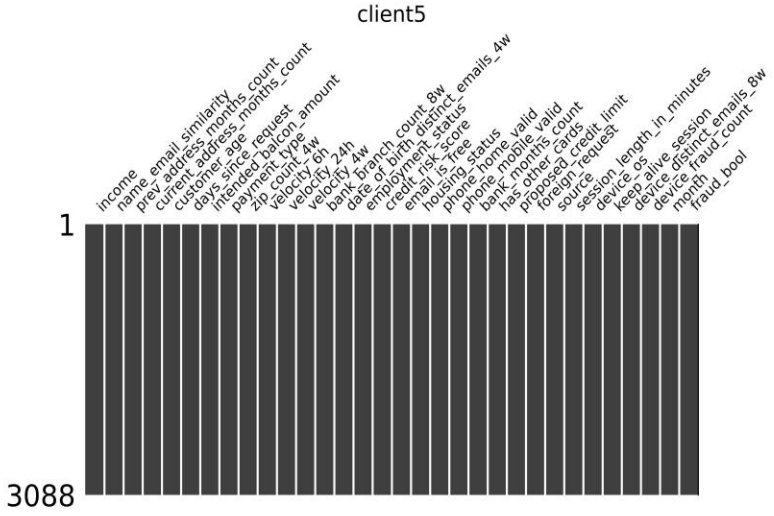


< Dataframe with manual missing values >

Method

● Client configuration





Missing Value columns

- Client1 : email is free
- Client2 : name email similarity, has other count, device fraud count
- Client3 : zip count 4w, velocity 6h, velocity 24h, velocity 4w, housing status
- Client4 : phone home valid, keep alive session
- Client5 : 없음

Related work

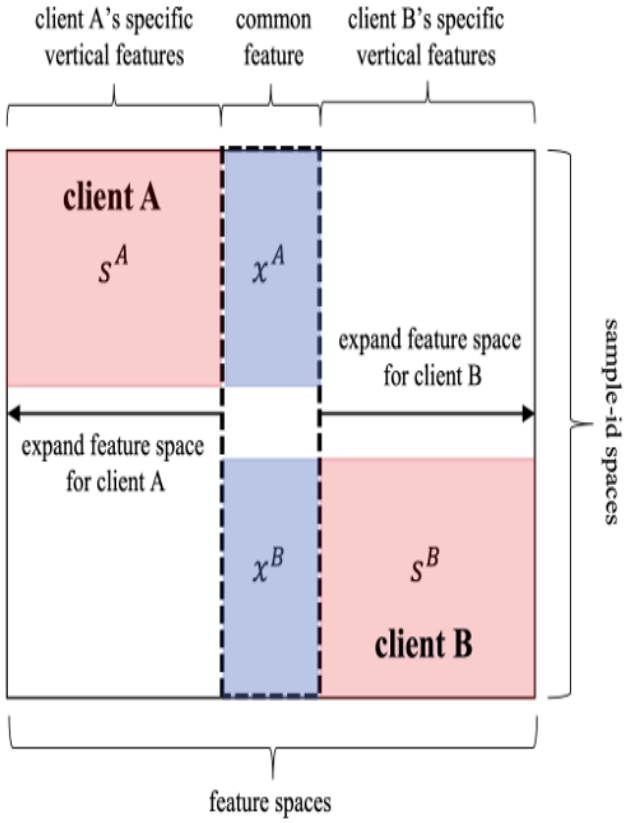


Fig1. Problem setting of personalized Progressive Federated Model

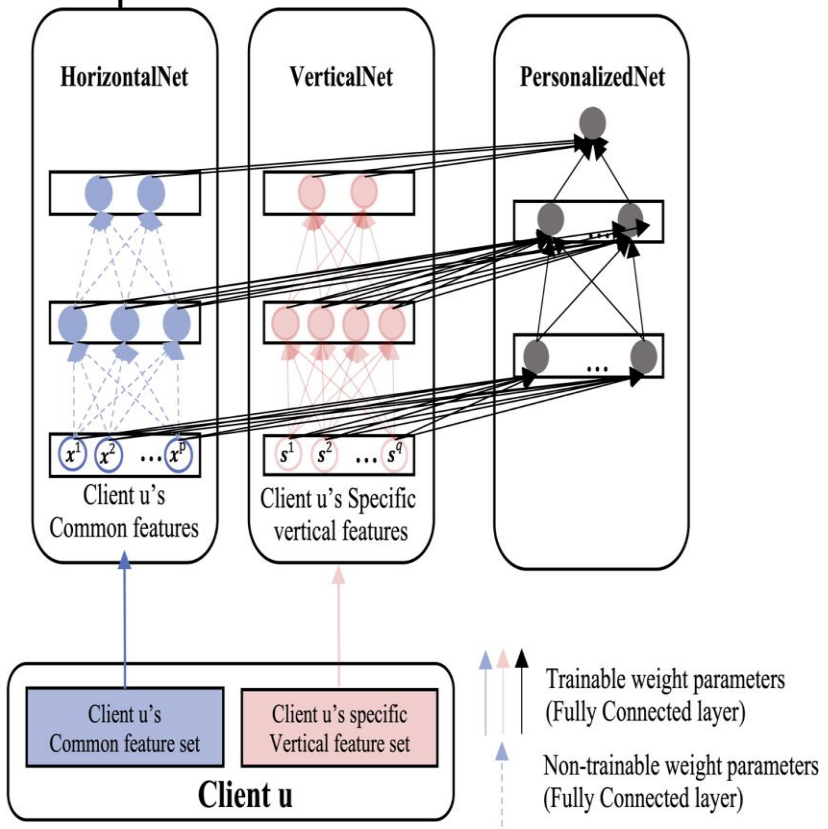


Fig2. Network architecture of the Personalized Progressive Federated Model

After HFL global model is trained by common features of all clients, each client has global weights

PersonalizedVertical(x^k, s^k)

Construct the PPFL model

$\mathcal{P} \leftarrow f(\omega^{c^{int}}, \omega^{v^{lat}}, \omega^{v^{int}}, \omega^{v^{lat}}, \omega^{p}; x^k, s^k)$

initialize $\omega^{c^{int}}$ to ω^{c}, and freeze the training of $\omega^{c^{int}}$

initialize $\omega^{c^{lat}}, \omega^{v^{int}}, \omega^{v^{lat}}, \omega^{p}$

for each personalization epoch e from 1 to E^p **do**

  for batch $b^p \in \mathcal{B}$ **do**

    $(\boldsymbol{\omega}_{e+1}^{c^{lat}}, \boldsymbol{\omega}_{e+1}^{v^{int}}, \boldsymbol{\omega}_{e+1}^{v^{lat}}, \boldsymbol{\omega}_{e+1}^{p})$

    $\leftarrow \boldsymbol{gradientdescent}(\boldsymbol{\omega}_{e}^{c^{lat}}, \boldsymbol{\omega}_{e}^{v^{int}}, \boldsymbol{\omega}_{e}^{v^{lat}}, \boldsymbol{\omega}_{e}^{p}), \ell, \eta^p; b^p)$

  end for

end for

return the PPFL model $\mathcal{P}$

김태현, et al. "Personalized Progressive Federated Learning with Leveraging Client-Specific Vertical Features: Model Development and Validation." 한국지능정보시스템학회 학술대회논문집 (2022): 260-260.

Related work (Algorithm)

< HFL – Fedavg >

```python
def fed_avg(client_list, total_round, epoch_per_round, sample_rate):
    num_sampled_clients = max(int(sample_rate * len(client_list)), 1)
    for r in range(1, total_round + 1):
        sampled_client_indices = sorted(
            np.random.choice(a=[i for i in range(len(client_list))],\
                             size=num_sampled_clients, replace=False).tolist())
        # Load global model to each client after the first round
        if r > 1:
            for idx in sampled_client_indices:
                client_list[idx].model.load_state_dict(global_model)
        # Local update
        for idx in tqdm(sampled_client_indices, desc='local_update'):
            for _ in range(epoch_per_round):
                client_list[idx].common_update()
        sampled_list = []
        for idx in sampled_client_indices:
            sampled_list.append(client_list[idx])
        # Global update
        global_model = global_avg(sampled_list)

    for cli in client_list:
        cli.model.load_state_dict(global_model)
```

< PPFL >

```python
def forward(self, X):
    hfl_hidden = X[:,self.common_idx]
    vfl_hidden = X[:,self.specific_idx]
    for i in range(self.last_layer_num+1):
        if i in self.activations_idx :
            ppfl_hidden = self.ppfl_model.layers[i](ppfl_hidden)
            hfl_hidden = self.horizontal_model.layers[i](hfl_hidden)
            vfl_hidden = self.vertical_model.layers[i](vfl_hidden)
        else :
            if i==0:
                ppfl_hidden = torch.cat([hfl_hidden,vfl_hidden],dim=1)
            else:
                ppfl_hidden = torch.cat([hfl_hidden,vfl_hidden,ppfl_hidden],dim=1)
            ppfl_hidden = self.ppfl_model.layers[i](ppfl_hidden)
            hfl_hidden = self.horizontal_model.layers[i](hfl_hidden)
            vfl_hidden = self.vertical_model.layers[i](vfl_hidden)

    return ppfl_hidden
```

- 모든 MLP는 4개 층과 ReLU 활성화 함수만 사용

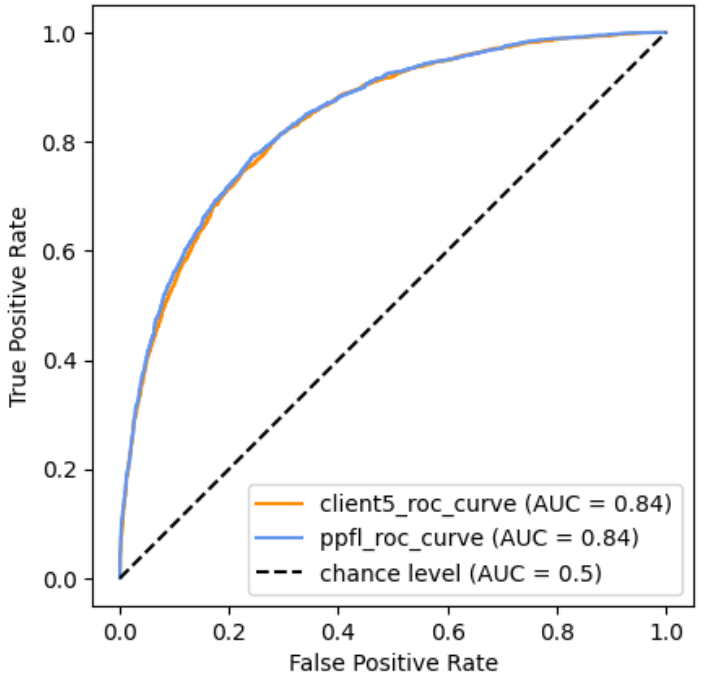
Outcome (Expected)

● Performance comparison

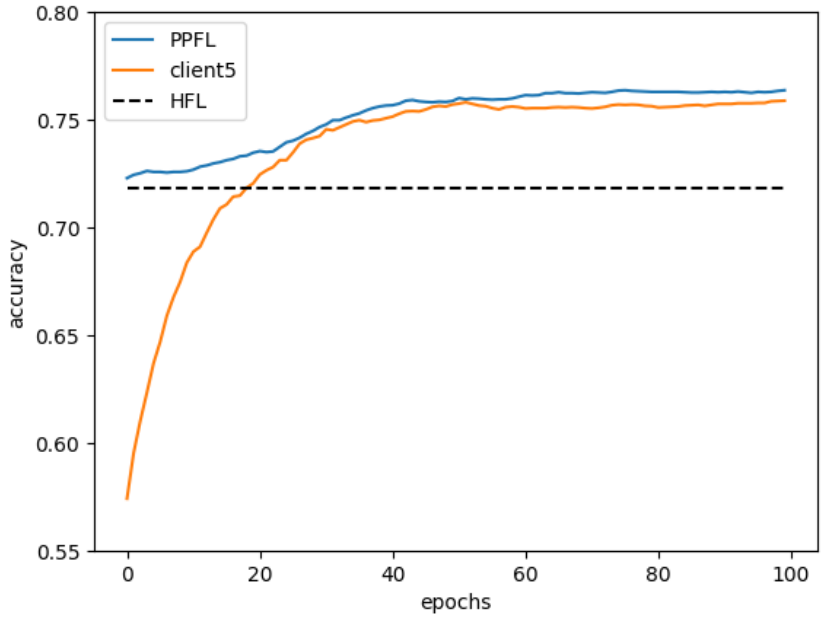
<Setting experimental>

* Case1: Client 5 + Client 1,2,3,4 (PPFL)
* Case2 : Client 5 (missing value none)
* Case3: Missing value imputation (mean, median etc.,)

ROC curve



Accuracy



< Results >

	Case1	Case2	Case3
Accuracy	0.7637	0.7588	0.7789
F1 score	0.7632	0.758	0.7776
Auroc	0.8428	0.8389	0.8634

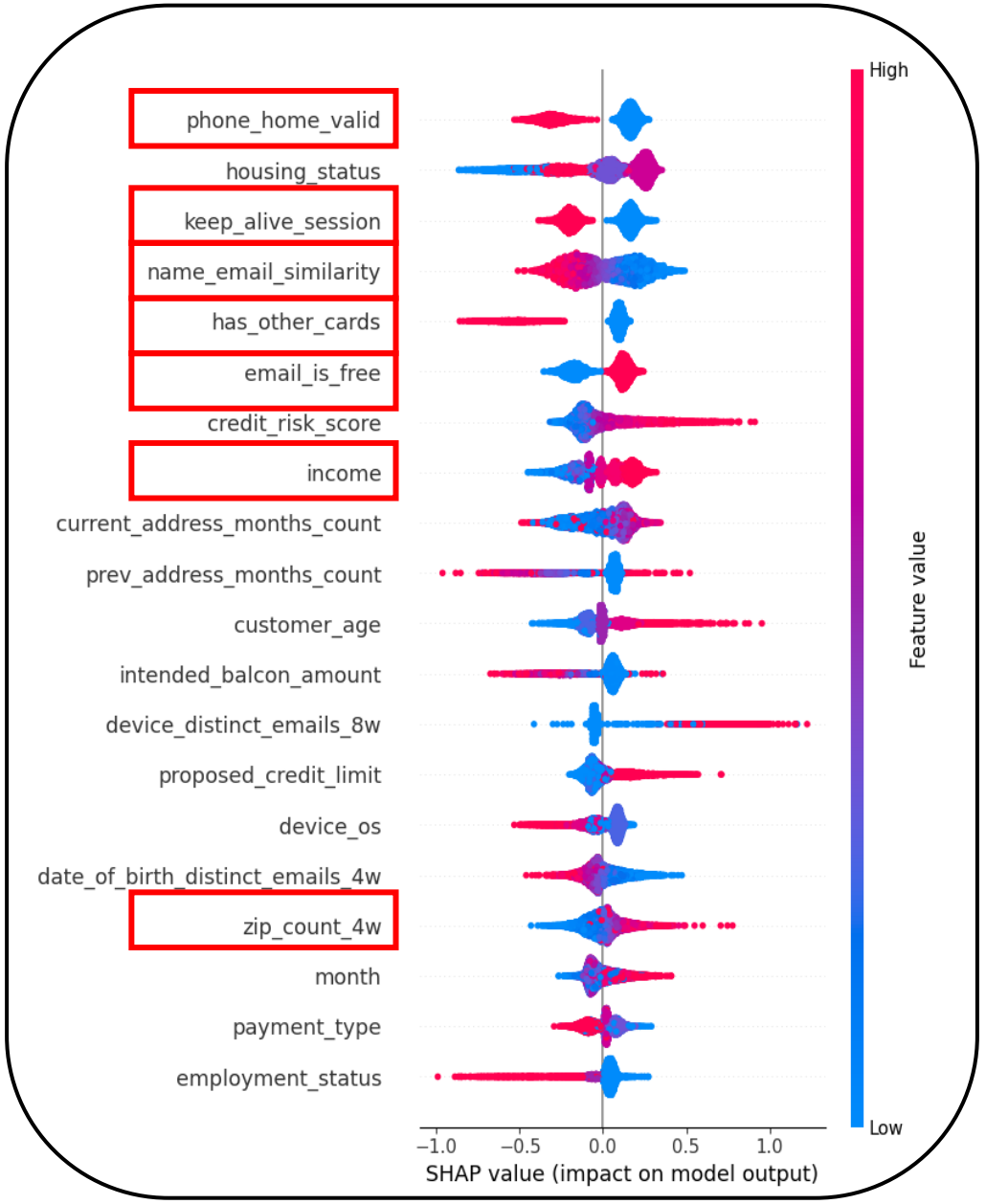
Outcome (Expected)

● Performance comparison

< Results >

	Case1	Case2	Case3
Accuracy	0.7637	0.7588	0.7789
F1 score	0.7632	0.758	0.7776
Auroc	0.8428	0.8389	0.8634

- 임의로 imputation한 case3의 성능이 제일 높음
- 학습 n수가 5배 차이나기 때문인것으로 보임
- Shap value 분석 결과 missing이 포함된 변수들이 성능 예측에 큰 영향을 미치는 것을 확인
  → 결과신뢰도와 해석에 영향을 미칠 수 있음



< SHAP Value >

Expectation & Limitation

- 결측치가 일정한 패턴 없이 무작위로 발생하는 경우 이 방법을 사용하는 것은 어려울 수 있음

- 그러나, 결측치가 일정한 패턴으로 생성되는 데이터에 대해서 적용하여 활용 가능함
  - 특정 시점 이후에 추가된 설문조사 항목으로 인해 생기는 결측치
  - 기계 고장으로 인해 특정 센서 데이터가 완전히 누락된 경우

- 우리의 연구 방법은 결과 분석에 대한 해석의 신뢰도를 향상시킬 수 있음

Team member roles

- 이주현*: 리더/프로젝트 매니저
- 최영조: 설계 및 개발 담당
- 양희민: 연구 및 분석 담당
- 김종현: 문서화 및 분석 담당

*=팀장

YONSEI UNIVERSITY
COLLEGE OF MEDICINE