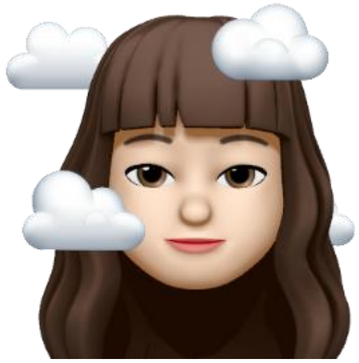
보이스피싱 탐지 알고리즘

시켜줘, 보아즈 명예경찰관

김다혜, 김성우, 김재민, 반소희, 홍주리

시켜줘, 보아즈 명예 경찰관



18기 분석 김다혜

한국외국어대학교
통계학과

18기 분석 김성우

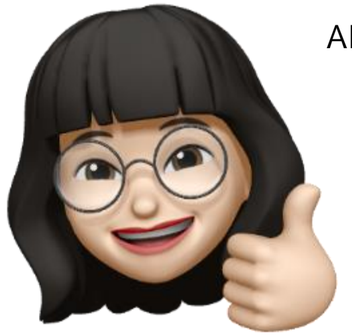
명지대학교
산업경영공학과

18기 엔지니어링 김재민

국민대학교
AI빅데이터융합경영학과

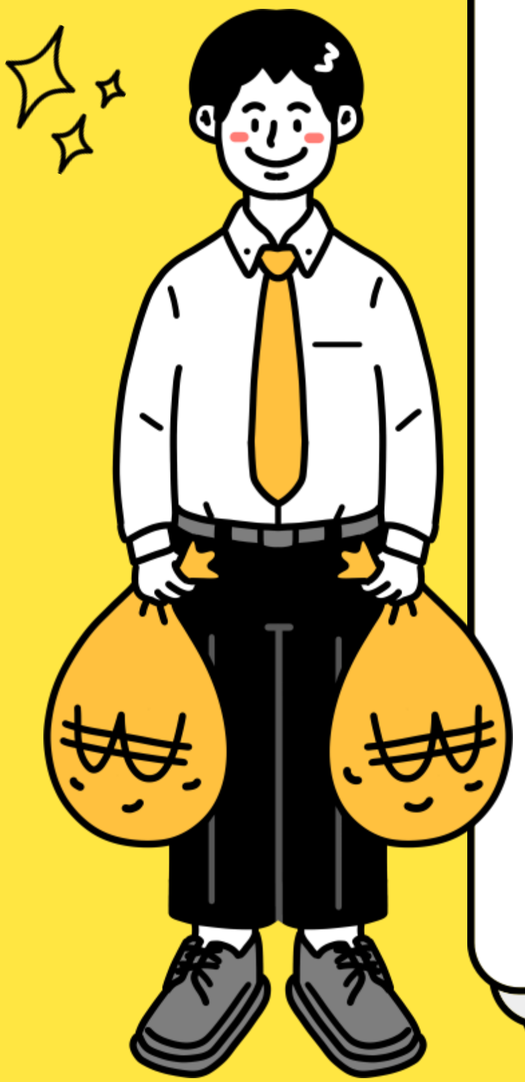
18기 분석 반소희

이화여자대학교
휴먼기계바이오공학부



18기 시각화 홍주리

숙명여자대학교
통계학과

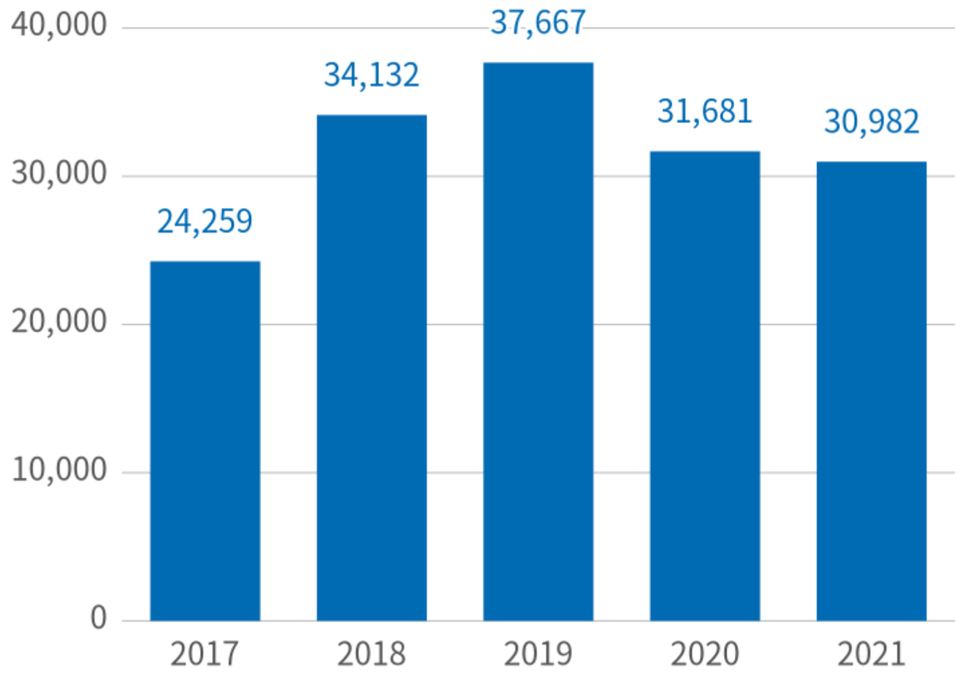


목차

보이스피싱을 당했거나,
위협에 노출되어본 경험이 있으신가요?

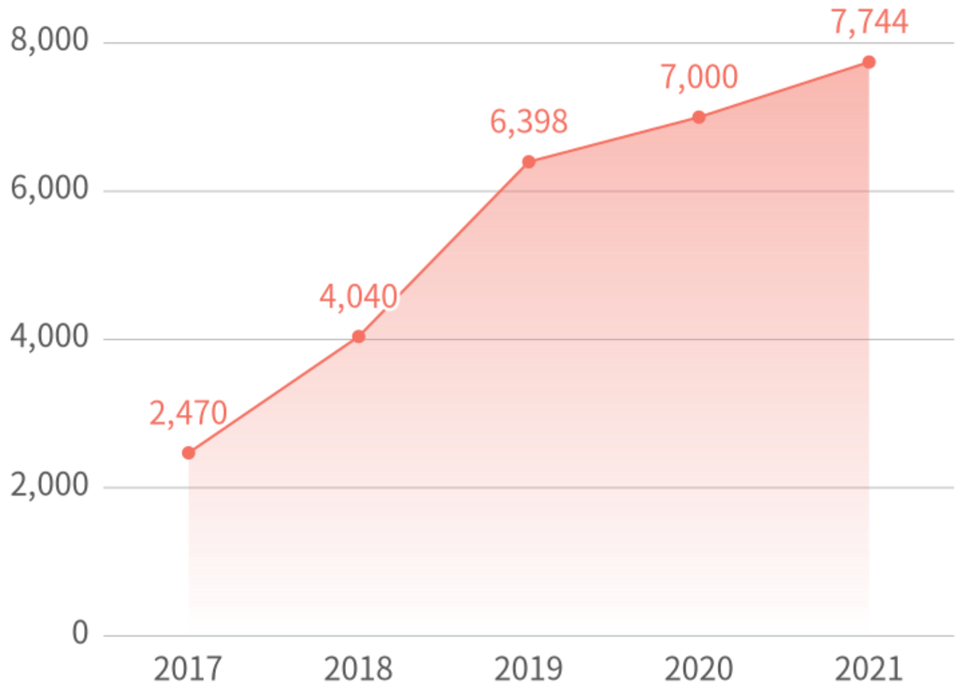
최근 5년간 보이스피싱 발생현황

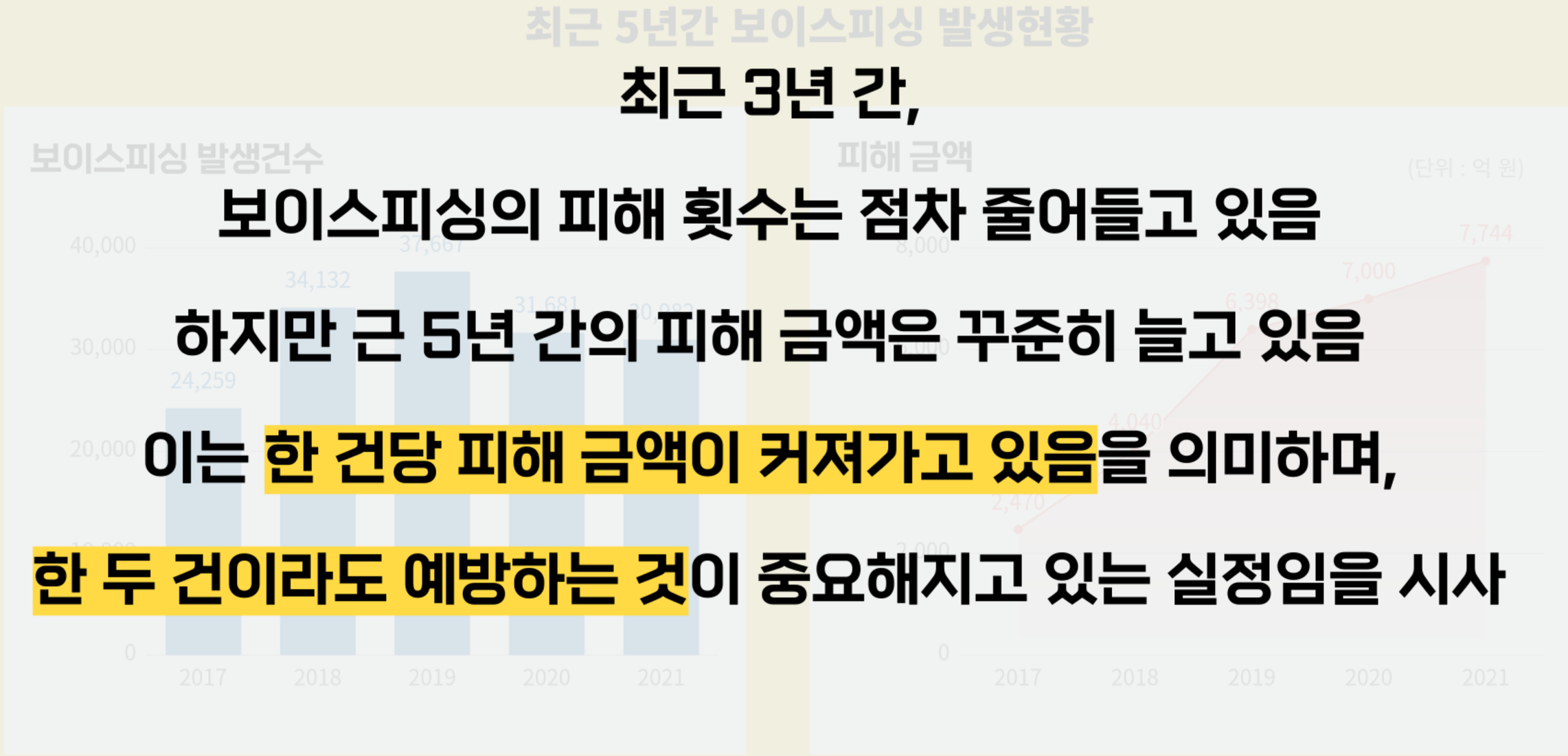
보이스피싱 발생건수



피해 금액

(단위 : 억 원)





최근 3년 간,

보이스피싱의 피해 횟수는 점차 줄어들고 있음

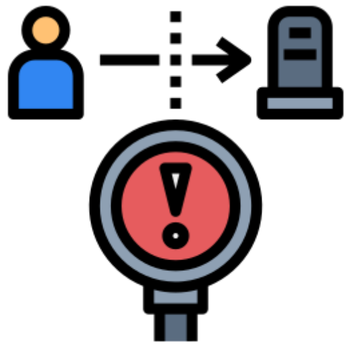
하지만 근 5년 간의 피해 금액은 꾸준히 늘고 있음

이는 한 건당 피해 금액이 커져가고 있음을 의미하며,

한 두 건이라도 예방하는 것이 중요해지고 있는 실정임을 시사

그래서 우리는,

모델 1



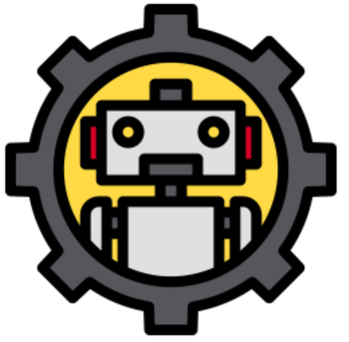
Real-Time 보이스피싱 탐지 모델

모델 2

통화 종료 후, 보이스피싱 최종 탐지 모델

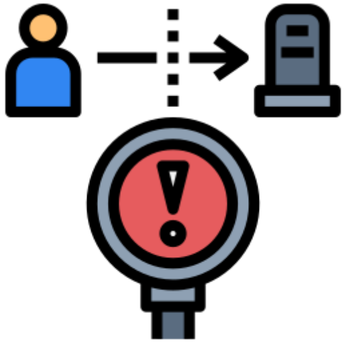
그래서 우리는,

음성 데이터



STT
(Speech To Text)

텍스트 데이터



Real-Time
보이스피싱 탐지 모델

통화 종료 후,
보이스피싱 탐지 모델

데이터 소싱

DATA

보이스피싱 (lable = 0)

- 그놈 목소리 (출처 : 금융감독원)
- 유튜브 보이스피싱 영상

일반 통화 (lable = 1)

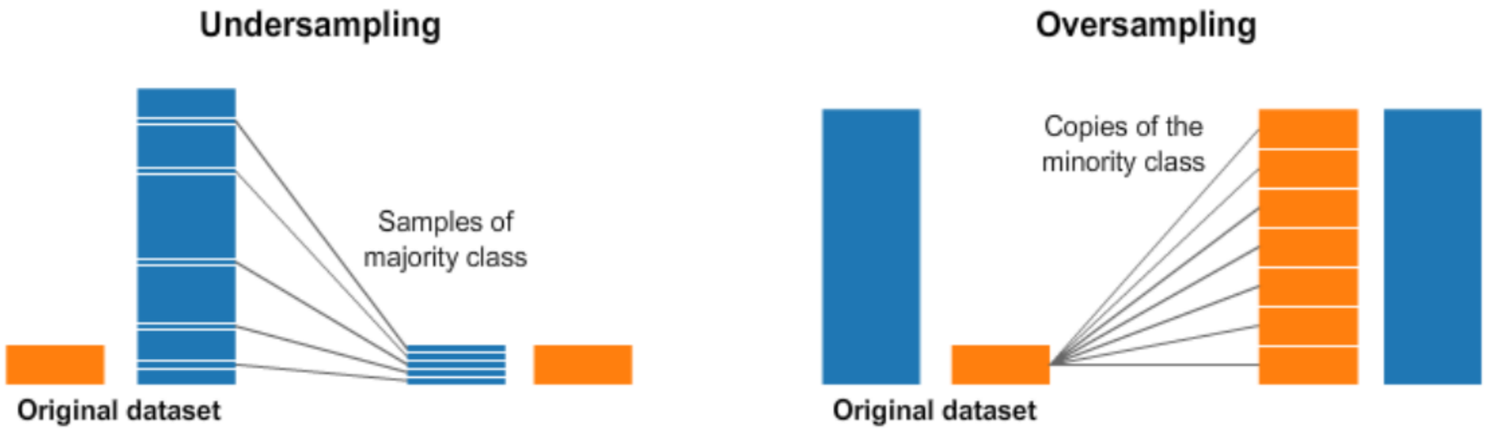
- AiHub 통화 데이터셋

보이스피싱 : 일반 통화 = 564건 : 55310건 = 1 : 100

"클래스 불균형"

1. Sampling
2. SMOTE
3. Text Augmentation

1. Sampling

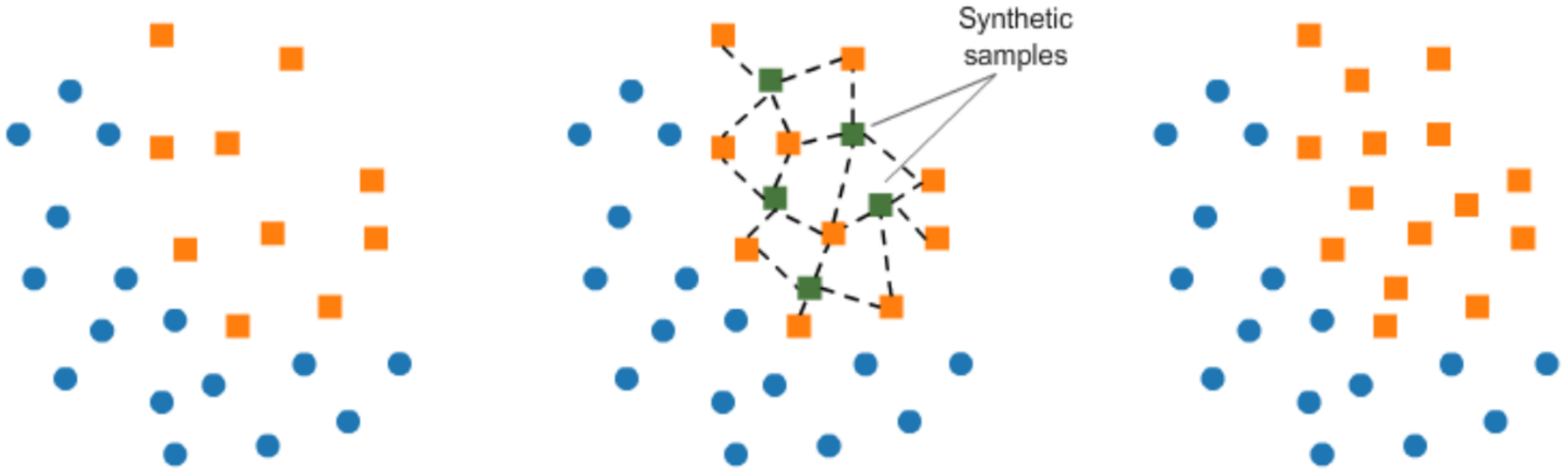


Sampling에는 Undersampling과 Oversampling이 존재

- Undersampling은 학습 데이터 셋이 크게 감소
- Oversampling은 과적합의 우려가 있음

해당 프로젝트에서 적절하지 않은 해결방법

2. SMOTE

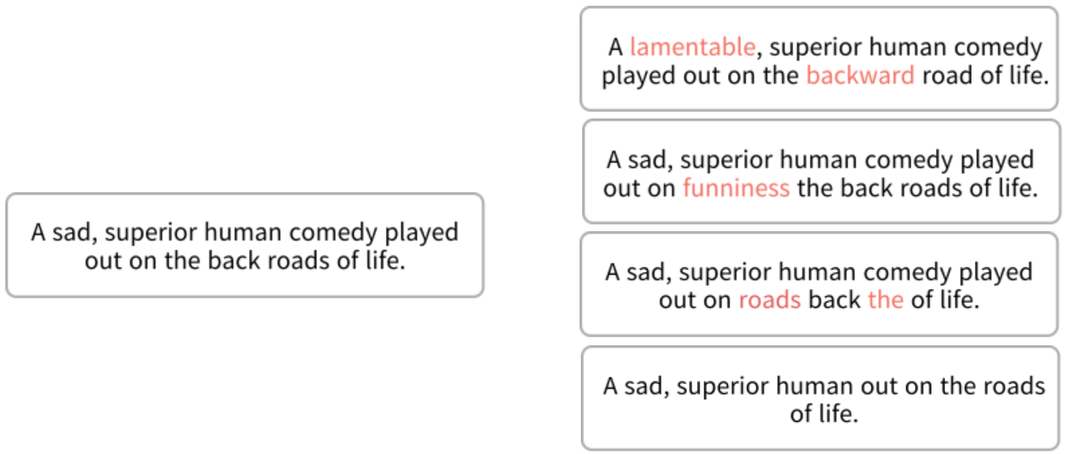


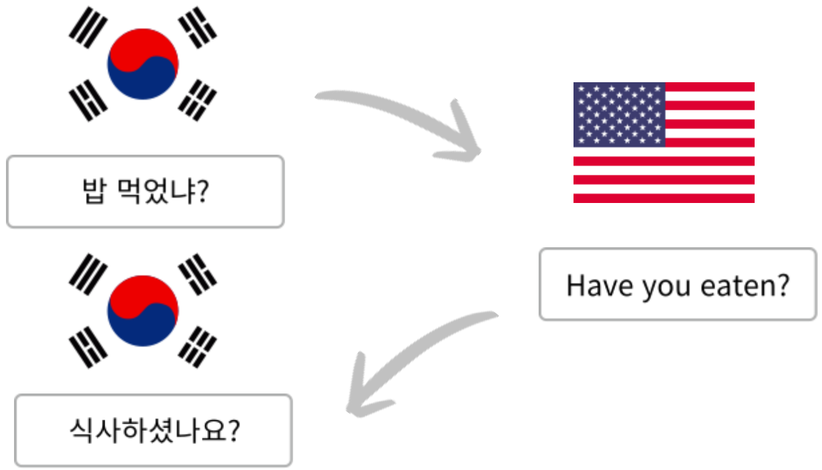
- 오버샘플링의 일종이지만, 과적합 문제를 해결하기 위해 만들어진 알고리즘
- 소수 클래스의 subset을 뽑아내어 새로운 데이터 생성
- 하지만, 벡터화가 되기 전인 text 데이터에선 활용이 불가한 기법

해당 프로젝트에서 적절하지 않은 해결방법

3.TEXT Augmentation

소수 클래스의 Data를 증강시키는 기법으로,
Back Translation과 Easy Data Augmentation이 있음





[Easy Data Augmentation]

- 2019년 EMLP에서 발표된 Text Data Augmentation 기법
- 평균 3%의 성능 향상을 기대할 수 있음

[Back Translation]

- 기존 Text를 다른 언어로 번역한 후, 다시 기존 언어로 번역
- NLP 분야에서 자주 활용되는 Augmentation 기법

해당 프로젝트에서 적절한 해결방법

활용한 Augmentation 기법

기법	내용
SR(Synonym Replacement)	특정 단어를 유의어로 교체
RI(Random Insertion)	임의의 단어를 삽입
RS(Random Swap)	문장 내 임의의 두 단어의 위치를 바꿈
RD(Random Deletion)	임의의 단어를 삭제
BT_JP(Back translation Japan)	한국어 → 일본어 → 한국어
BT_EN(Back translation English)	한국어 → 영어 → 한국어

모델에 따라 알맞은 Augmentation 기법이 존재하지 않을까?

Text Augmentation과 모델 Select을 동시에 진행

[Model 1]

Real-Time 보이스 피싱 탐지 모델

전반적인 모델 설계 과정

1

음성 데이터

2

VITO

STT
(Speech To Text)

3

텍스트 데이터셋

K-Fold 후 최적 모델 선정

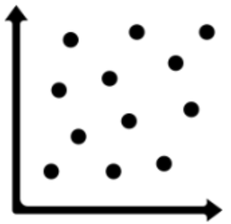
4

Data Augmentation
- Easy Data Augmentation
- Back Translation

5

데이터 전처리

6



벡터화(Tf-idf)

7

모델 학습
- LGBM
- AdaBoost
- XGB
- Random Forest

STT(Speech To Text)

- 앞서 수집한 음성 데이터를 VITO STT를 활용하여 text 파일로 변환
- VITO는 STT 기술 기반의 '소머즈(Sommers)' 엔진 적용
  - 한국어 구어체, 자유 발화, 소음 등의 환경에 노출된 통화 음성인식에 특화된 엔진
  - 욕설, 간투어 필터링 기능 등을 제공
    - ■하지만 보이스피싱 특성 상, 욕설 및 간투어가 분류에 영향을 끼칠 수 있으므로 제거하지 않음

Text Augmentation

기법	내용
SR(Synonym Replacement)	특정 단어를 유의어로 교체
RI(Random Insertion)	임의의 단어를 삽입
RS(Random Swap)	문장 내 임의의 두 단어의 위치를 바꿈
RD(Random Deletion)	임의의 단어를 삭제
BT_JP(Back translation Japan)	한국어 → 일본어 → 한국어
BT_EN(Back translation English)	한국어 → 영어 → 한국어

RS는 활용하지 않음

- ML모델은 단어의 순서를 고려하지 않음
- 즉, 순서를 바꾸는 RS의 경우 Augmentation이 무효할 것이라 판단하여 제거

Text Augmentation

0개 선택

1개의 데이터셋 생성

None

1개씩 선택

5C1 = 5이므로,
5개의 데이터셋 생성

SR

RI

⋮

2개씩 선택

5C2 = 10이므로,
10개의 데이터셋 생성

SR + RI

SR + RD

⋮

3개씩 선택

5C3 = 10이므로,
10개의 데이터셋 생성

SR + RI + RD

SR + RI + BT_EN

⋮

4개씩 선택

5C4 = 5이므로,
5개의 데이터셋 생성

SR + RI + RD
+ BT_JP

SR + RI + BT_EN
+ BT_JP

⋮

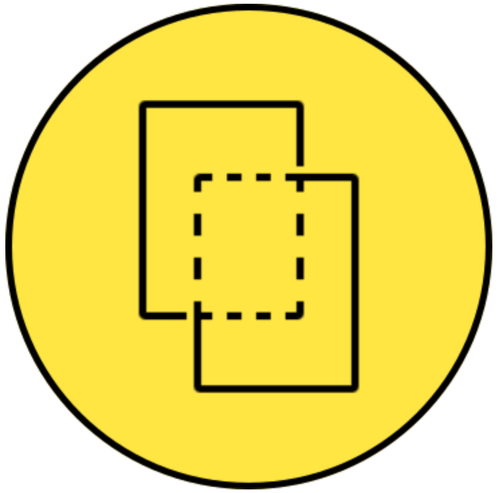
5개 모두 선택

1개의 데이터셋 생성

SR + RI + RD+
BT_JP + BT_EN

각 모델별로 최적의 데이터셋을 탐색하기 위해
1 + 5 + 10 + 10 + 5 + 1 = 32
총 32개의 데이터셋 생성

데이터 전처리



중복 & null값 제거

텍스트 cleansing

- 한글이 아닌 문자 제거
- 불용어 제거

벡터화

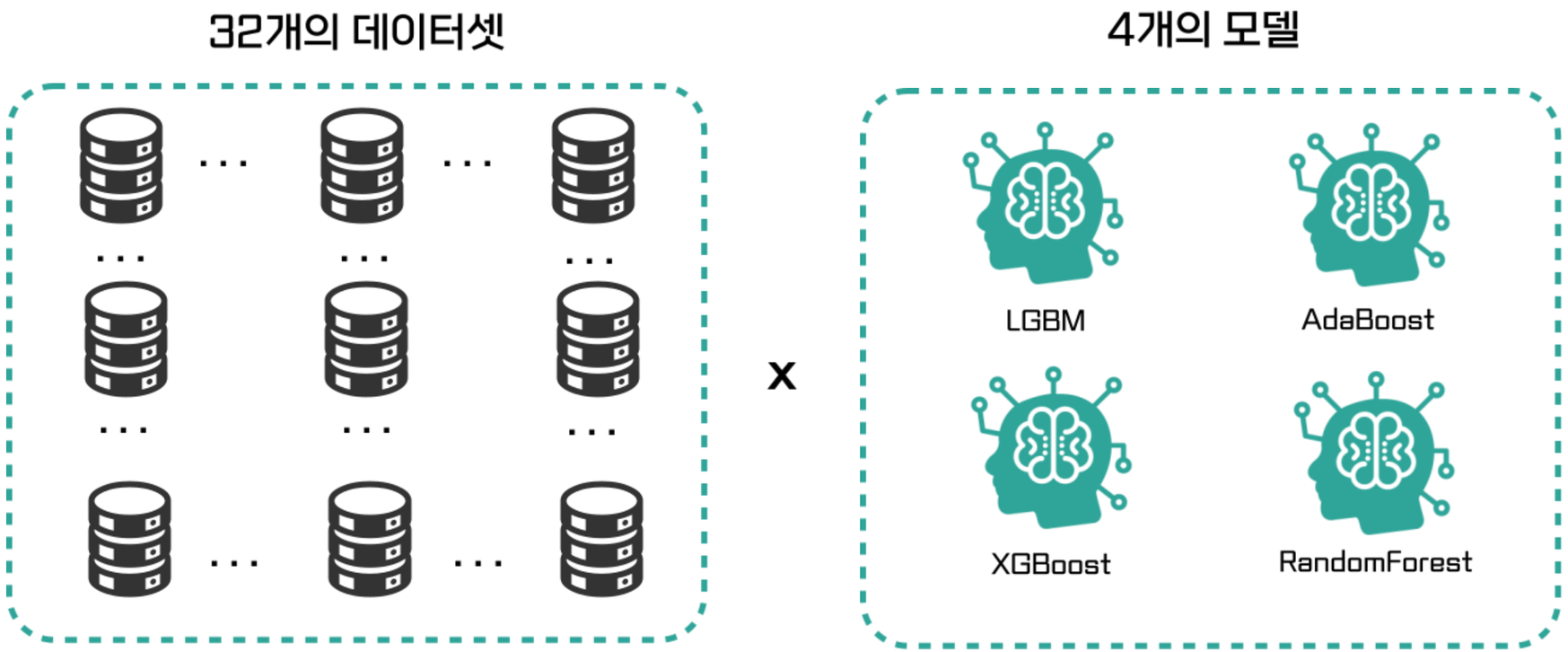
Tf-idf

은행, 계좌 등 보이스피싱과 관련된
단어에 더 큰 가중치가 반영됨

Countvectorizer

안녕, 아니 등 일반적인 단어에
더 큰 가중치가 반영됨

모델 학습



- 앞서 선정한 32개의 데이터셋에 4개의 모델을 학습
- 즉, 32 x 4 = 128번의 실험을 통해 각 모델별 최적의 증강 데이터셋을 탐색
  - 5-fold validation의 평균 F1 score 비교

모델 학습

F1-score	None	RD	SR_RI	SR_RD	...	ALL
RF	0.863	0.888	0.886	0.894	...	0.898
Adaboost	0.966	0.970	0.974	0.966	...	0.954
XGB	0.966	0.972	0.965	0.966	...	0.966
LGBM	0.958	0.972	0.971	0.973	...	0.969

HyperParameter 튜닝

AdaBoost with SR&RI
F1 score: 0.97408

XGB with RD
F1 score: 0.9723

LGBM with SR_RD
F1 score: 0.97345

Optuna

Tuned AdaBoost with SR&RI
F1 score:0.97488(0.001 ↑)

Tuned XGB with RD
F1 score: 0.982(0.01 ↑)

Tuned LGBM with SR_RD
F1 score: 0.98352(0.01 ↑)

최종 선정한 모델

최종 선정된 모델의 작동 과정

[Model 2]

통화 종료 후,
보이스피싱 최종 탐지 모델

전반적인 모델 설계 과정

1

음성 데이터

2

VITO

STT
(Speech To Text)

3

텍스트 데이터셋

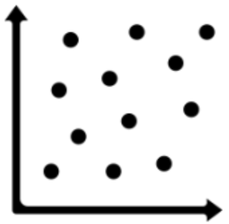
4

Data Augmentation
- Easy Data Augmentation
  - ○ 모델1에서 빠졌던 RS 추가
- Back Translation

5

데이터 전처리

6



word embedding

7

모델 학습
- LSTM

Text Augmentation

기법	내용
SR(Synonym Replacement)	특정 단어를 유의어로 교체
RI(Random Insertion)	임의의 단어를 삽입
RS(Random Swap)	문장 내 임의의 두 단어의 위치를 바꿈
RD(Random Deletion)	임의의 단어를 삭제
BT_JP(Back translation Japan)	한국어 → 일본어 → 한국어
BT_EN(Back translation English)	한국어 → 영어 → 한국어

- RS도 포함하여 모든 Augmentation 기법을 활용하여 학습
- LSTM은 sequential 모델이기 때문에 순서를 바꾸는 기법도 유효
- 또한 신경망 모델이므로, 최대한 많은 데이터를 확보

LSTM 모델 설계

LSTM

- weight balancing
- binary cross entropy
- adam optimizer

LSTM 모델 설계

x

LSTM

- weight balancing
- binary cross entropy
- adam optimizer

SR+RI+RD+RS+
BT_JP+BT_EN

F1 score : 0.9639

LSTM 모델 설계

x

LSTM

- weight balancing
- binary cross entropy
- adam optimizer

SR+RI+RD+RS+
BT_JP+BT_EN

F1 score : 0.9639

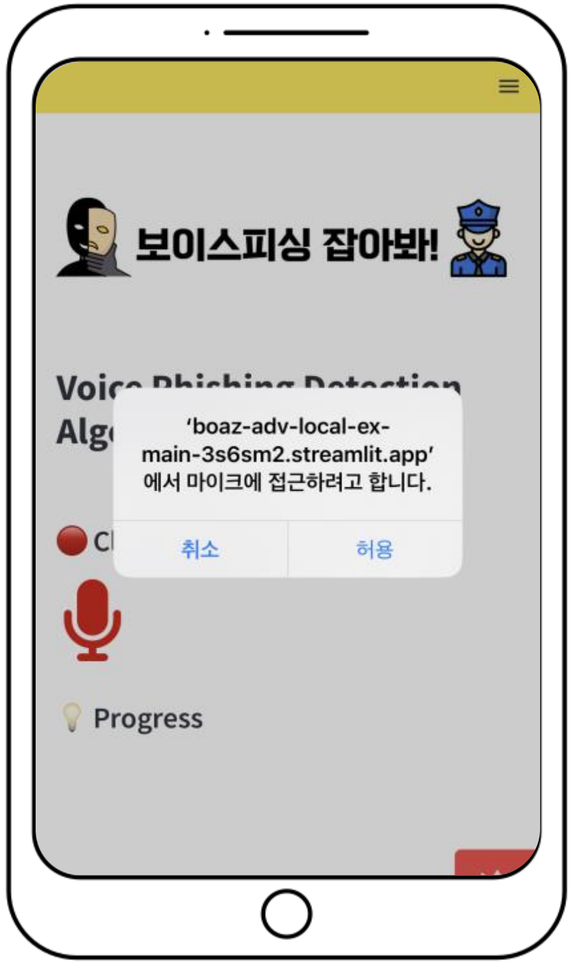
x

LGBM

SR + RD

F1 score : 0.9835

웹 배포



Streamlit

- 파이썬 기반의 웹프레임워크
- 머신러닝, 딥러닝 모델이나 데이터 시각화를 웹서비스로 쉽게 배포 가능

웹 배포


보이스피싱 잡아봐!
Progress


보이스피싱 잡아봐!
Voice Phishing Detection Algorithm
Click to record
Progress
Stop Recording
Speech To Text 진행 중...
Call Classification Model & Encoder
Finish
결과보기
위험
보이스피싱 확률이 94.8% 입니다.
0:00 / 2:45
RESULT TEXT
직원인지 점심 도시락 안 싸 님들 알면 안됨까요 황수조 예 강상수 강상수라고요 예 농협이면 한 7년정도 분명히 죄송합니 왜요 예 저희가 이 본이 강상수님 중심으로 한 금융 보험 설계사님 건들있습니다 잠시 발랭인 대포통장한거야 신분증 농협은행과 신한은행 통장이 발견 되면데 시중승들에 대해서하신 내용 있으실까요 장소가 아니고 함소 안해요 예 장상씨요 상승 아니야 장난이야 장난 아니야 참 장수 소갈비 상수 우선은 잘 모르겠는데 여기는 왜가 예 주임님 건건이 강상 순찰 안 터질지 모른다는 점이조 39살인데 어 형수는 잘 모르겠어 하는 스리가 안 되니까 예 저희가 합선이 선정적인 농협은행이랑 신한은행 발급 네 병원 쪽에 보니까요 네 2020년 6월 1등이요 그럼십시오 신한은행은 신규 동주 집에서 발견이 되었거든요 공인인증서 발급 다 주신 게 맞으십니까 저는 발급을 안했는데 씨라면 어디서 하는데요 예 주식회사 잘나가는 신경 안 들여 주십시오 예 제가 오늘 범진한테 연락 드린 이유는 임차인이 대포 통장 개설돼요 사실 확인차 연락드린거고요 예 17 다시 범죄 살림되어서 시세가 발생되는 부분인데 저희 주유소에서 전반적인 내용을 주 새해 공차일 거야 거실 소파도 없으시고 신분증 땃네 미리 이준 사항으로 연락드린 겁니다 본인 명의의 통장이 발견 됐다고 그래서 저희가 무조건 다 해주는거 있지 않나요 어느 정도는 피의자 보고있는데잘아 아직까지 죄생 진료를 준거잖아요 그래서 저희가 뭔색 친구들 도와주고 하는 부분이 있고요 제 지금이네 아신다 모르시입니다 간단히 주시면 되겠습니다 그럼 지금 죄송하지만 지금 전화 거신 분 어디시라고요 예 저는 서울중앙지검 지역 국립 전당진 7074입니다 김용재 수사관님요 예 수리는 어떻게 감당하기 더 빌리시면 됩니다 예
Voice Phishing Probabilty
Text Length

한계점

- 다양한 기법을 적용했음에도 존재하는 class imbalance 문제
- 리소스 부족으로 더 많은 신경망 모델을 실험하지 못한 점
- 실제 제품에 실시간 처리 파이프라인을 적용시키지 못한 점

감사합니다!