【붙임 2】

데이터 분석 최종결과보고서

Ⅰ. 참가자 정보

제 목	의사결정 트리 기반의 보이스피싱 데이터 분석 및 경찰력 선제적 분배	
팀 명	K-CLUE	
성 명	장지윤	
연락처	휴대폰	010-3193-0574
	E-mail	fick17@korea.ac.kr

Ⅱ. 개요

○ (분석/시각화) 목적

1. 보이스피싱 신고 데이터를 분석 및 시각화하여, 범죄 경향 및 추이를 파악하고 대응에 도움이 되도록 한다.
2. 범죄 유형을 예측하는 모델을 통해, 보이스피싱 사건임을 선제 파악하여 대응 속도를 높인다.
3. 피해자 신고 시 위도, 경도를 예측하는 모델로 구역별 월간(연간) 범죄 발생 건수를 예측하고, 경찰 인력 충원 및 감소를 결정한다. 경찰력 분배를 최대 효율로 이루는 것을 목표로 한다.

○ 배경 및 필요성

인터넷과 모바일 기술의 발전으로 사기 우려가 증가하면서, 보이스피싱이 대한민국 국민들에게 큰 문제가 되고 있다. 경찰과 각종 기관이 보이스피싱 문제에 대한 각종 대처와 대책을 제시하고 있지만, 아직도 많은 국민들이 보이스피싱의 피해를 당하고 있다. 보이스피싱 범죄 피해는 2018년부터 2021년까지 점차 증가하였으며, 2022년에는 소폭 감소하였다. 하지만 2022년도에도 여전히 5147억의 피해 금액 및 2.05만명의 피해자가 존재하며, 갈수록 지능형 범죄 역시 늘어나고 있다.

또한, 경찰력 분배를 통한 112신고에 대한 차별적 경찰 대응은 필수적이다. 지난 10년간 112신고 및 경찰인력의 추세를 살펴보면, 2007년 약 620만 건의 112신고가 접수된 이후 10년간 112신고접수건수는 대폭 증가하여 2016년에는 약 1,960만 건의 3배 이상이 증가한 반면, 경찰인력은 2007년 96,324명에서 2016년 114,658명으로 약 19% 증가하여 경찰인력의 증가가 치안수요의 증가를 따라가지 못하였다. 이러한 경찰력의 한계에 대하여 이전부터 지속적으로 문제제기가 이루어져왔고, 그리하여 112신고 중 긴급한 신고에는 신속하게 대응하는 반면, 긴급하지 않은 신고에 대해서는 시간을 두고 적절하게 대응하는 차별적인 경찰대응방식을 적용할 필요가 있다는 주장이 제기되었다. 차별적인 경찰대응의 효과에 대해서 전반적으로 긍정적인 분석결과들이 보고되었다. 구체적으로 차별적 경찰대응 실시 후 긴급하지 않은 신고에 대한 비출동 조치로 인해 신고출동건수는 감소하였으며, 경찰의 112 긴급신고에 대한 평균도착 시간이 2010년 4분 41초에서 2012년에는 3분 52초로 약 50초 가까이 단축된 것으로 나타났다(이강훈, 2014: 370).[1)]

지능형 보이스피싱 범죄를 근절하고 경찰력 효율적 분배를 위해, 데이터적 관점의 접근으로 선제적 예측 및 대응을 제안한다. 여러 측면의 범죄 경향 및 추이를 파악하고, 기계학습 기법을 사용해 선제 예측을 실현하려 한다.

1) 조준택, 김강일, 박현호, 범죄예방을 위한 112신고 자료의 활용방안 연구(한국치안행정논집 제15권 제1호), p6, 2018

○ 분석/시각화 결과 내용 요약

1. 데이터 분석

주어진 보이스피싱 데이터의 '발생 시간', '경찰청', '요일'에 대하여 유의미한 분석 결과가 있었다. '발생 시간'의 경우 00시-7시 경의 발생률이 현저히 낮았고, 점차 발생률이 상승해 12-14시 경에 가장 높은 발생률을 보였다. '경찰청'의 경우 19(충남), 13(대전), 31(세종) 순으로 발생률이 높은 것을 알 수 있었다. '요일'의 경우 0~4(평일)의 발생 횟수가 5~6(주말)보다 확연히 높은 것이 확인되었다. 이는 전체적인 범죄의 경향성을 벗어난, 보이스피싱 범죄만의 특징이다.

2. 모델

2-1. 보이스피싱 여부 예측 모델

특정 신고가 보이스피싱 범죄에 대한 건인지 아닌지를 판별하는 모델을 추출했다. 데이터의 다른 정보들을 이용해 범죄 구분을 예측하는 decision tree 기반의 분류기 모델이다. 성능 지표는 f1-score를 채택하였으며 0.994라는 우수한 지표를 기록하였다. 범죄 당시의 정보를 통해 범죄 여부를 알아낼 수 있다는 사실이 고무적이다. 이를 활용하면 범죄 사실에 무지한 피해자에게 적절한 대응 방안을 마련해줄 수 있을 것으로 보인다.

2-2. 위도, 경도 예측 모델

보이스피싱 범죄의 위도, 경도를 예측하는 모델을 추출했다. 위에서 기술한 모델과 마찬가지로 데이터의 다른 특성들을 통해 위도, 경도를 각각 독립적으로 예측하는 원리이며, Gradient Boosting Regressor 방법을 채용하였다. 결측치를 처리하지 않고 측정했을 시 경도는 0.0971, 위도는 0.0674의 평균 제곱 오차를 띠었다. 해당 모델은 실제 신고 상황에서 피해자가 위치 정보를 전달할 수 없을 때 활용할 수 있을 것으로 보인다. 또한 해당 모델을 통해 앞으로의 보이스피싱 범죄 발생 건수를 구역별로 예측해보았다. 이 예측치에 따라 인력을 충원하거나 감원하는 등, 효율적인 경찰력 배분에 활용할 수 있을 것이다.

Ⅲ. 분석/시각화 결과 상세내용

○ 분석/시각화 결과 상세내용

1. 데이터 전처리

1-1. 데이터 병합

주어진 KP2020, KP2021, NPA2020 세 데이터를 모두 하나의 형식으로 병합하여 모델에 사용할 train이라는 데이터를 제작하였다.

1-1-a. KP2020, KP2021 처리

KP2020, KP2021 데이터는 각각 77,077개, 2,594,060개의, 10가지 항목으로 구성된 119 신고 데이터이다. 다음은 각 항목에 대한 내용을 표로 나타낸 것이다.

항목	내용
RECV_DEPT_NM	접수부서 코드
RECV_CPLT_DM	접수완료일시
NPA_CL	경찰청구분
EVT_STAT_CD	사건상태코드
EVT_CL_CD	사건종별코드
RPTER_SEX	신고 성별 - 1 : 남자 2 : 여자 3 : 불상
HPPN_PNU_ADDR	발생지점(PNU)
HPPN_X	발생좌표X
HPPN_Y	발생좌표Y
SME_EVT_YN	동일사건여부

- KP 데이터 내, RECV_DEPT_NM은 NPA_CL 데이터가 나타낼 수 있었으며, HPPN_NPU_ADDR은 HPPN_X, HPPN_Y로 나타낼 수 있기에 RECV_DEPT_NM과 HPPN_NPU_ADDR 열을 제거하였다.
- RECV_CPLT_DM 데이터는 '연도/월/날짜/ 시간:분:초' 형식이다. 시간 데이터를 활용하기 위해, RECV_CPLT_DM 데이터 대신 YEAR, MONTH, DAY, HOUR 열을 추가하였다.
- KP 데이터 내 'RPTER_SEX' 열에는 1, 2, 3을 제외한 None이라는 결측치가 26,313개 존재한다. KP데이터와 NPA 데이터를 병합하기 위해서 'RPTER_SEX' 데이터에서 결측치를 제거하고 데이터 타입을 통일할 수 있도록 했다.
- SME_EVT_YN은 주어진 주제와 맞지 않는 열이기에 모델에 정확도를 높이고자 제거하였다.

1-1-b. NPA2020 처리

NPA2020 데이터는 1,178,244개의, 10가지 항목으로 구성된 119 신고 데이터이다. 다음은 각 항목에 대한 내용을 표로 나타낸 것이다.

항목	내용
RECV_DEPT_DT	접수완료일자
RECV_CPLT_TM	접수완료시간
NPA_CL	경찰청구분
EVT_STAT_CD	사건상태코드
EVT_CL_CD	사건종별코드
RPTER_SEX	신고 성별 - 1 : 남자 2 : 여자 3 : 불상
HPPN_OLD_ADD	발생구주소
HPPN_X	발생좌표X
HPPN_Y	발생좌표Y
SME_EVT_YN	동일사건여부

NPA2020 데이터를 KP 데이터와 병합하기 위해 KP 데이터 형식에 맞춰 가공하였다.

- 접수완료일자가 '20200101'등의 8자리로 표현되어 있다. 이를 연도(20), 월(01), 일(01)의 형태로 분할하였다.
- 접수완료시간의 경우에는 여섯 자리가 아닌 경우에 앞에 0을 붙여 여섯 자리로 만들었다. '시', '분', '초' 중에서 '시'만 활용할 것이므로 앞 두 자리만 분리하였다.
- '발생구주소'는 발생좌표 X,Y와 의미상으로 중복되기 때문에 제거하였다.
- 동일 사건 여부는 본 주제와 무관하기 때문에 제거하였다.
- RPTER_SEX 열의 ' '(공백)과 '{'라는 결측치가 존재하였다. 해당 항들을 제거해준 뒤 int형으로 변환하였다.
- HPPN_X, HPPN_Y의 결측치의 경우에는 0으로 값이 들어가 있었기 때문에 KP 데이터와 병합한 후 평균값으로 처리하기 위해 0을 np.nan 값으로 대체해주었다.

1-1-c. 데이터 병합 및 열 분석

KP2020, KP2021, NPA2020 데이터를 모두 병합하여 train 데이터를 제작하였다. 또한, 요일별 분석을 위해서 python datetime을 이용하여 WEEKEND(접수완료 시 해당 요일)을 항목에 추가하였다. train 데이터의 항목 및 내용은 다음과 같다.

항목(데이터 타입)	내용
NPA_CL(int64)	경찰청구분
EVT_STAT_CD(int64)	사건상태코드
EVT_CL_CD(int64)	사건종별코드
RPTER_SEX(int64)	신고 성별 - 1 : 남자 2 : 여자 3 : 불상
HPPN_X(float64)	발생좌표X
HPPN_Y(float64)	발생좌표Y
YEAR(int64)	접수완료시 해당 연도
MONTH(int64)	접수완료시 해당 월
DAY(int64)	접수완료시 해당 일
HOUR(int64)	접수완료시 해당 시
WEEKEND(int64)	접수완료시 해당 요일

1-2. 데이터 전처리

각 열마다 이상치 확인 및 처리, 결측치 확인 및 처리를 진행하였다.

1-2-a. 'NPA_CL'

데이터 내에 대전(NPA_CL = 13), 충남(NPA_CL = 19), 세종(NPA_CL = 31) 이외 지역 데이터 또한 29,987개 포함되어 있었다. 현재 데이터는 대전, 충남, 세종의 119 신고 데이터를 기반으로 분석하므로 이외 지역 신고 데이터는 모두 제거하였다. 결측치는 존재하지 않았다.

1-2-b. 'HPPN_X', 'HPPN_Y'

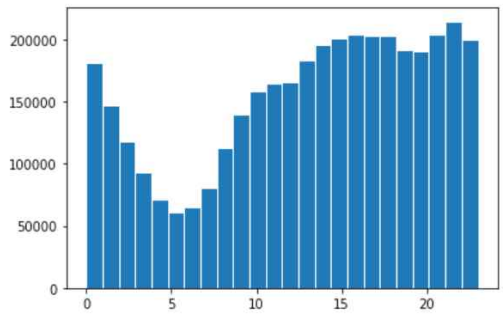
각각 928,814개의 결측치가 존재하였다. 이에 각 열의 평균값으로 결측치를 대체하는 작업을 수행하였다.

1-2-c. 'EVT_STAT_CD', 'EVT_CLCD', 'RPTER_SEX', 'YEAR', 'MONTH', 'DAY', 'HOUR', 'WEEKEND'

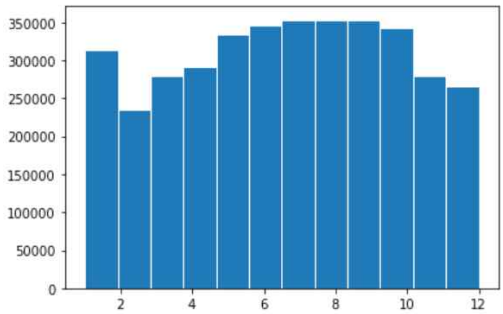
결측치 및 이상치가 존재하지 않았다.

2. 탐색적 데이터 분석

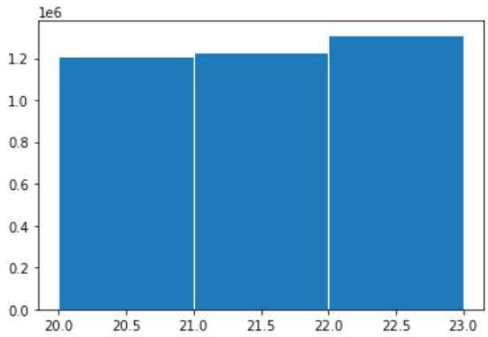
2-1. 전체 자료 분석 (세로축 : 발생 건수, 가로축 : 해당 항목별)



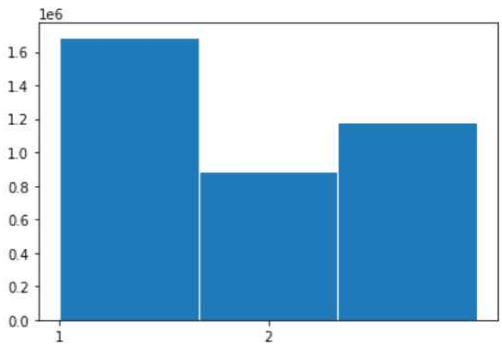
a. 시간별



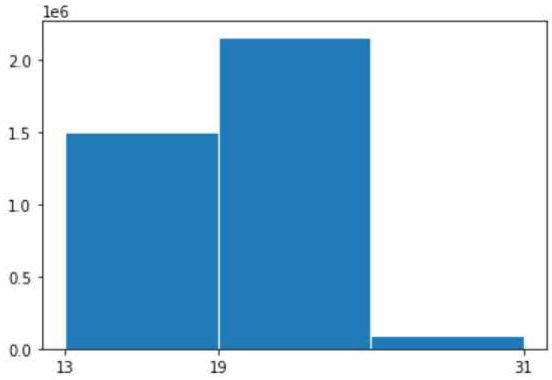
b. 월별



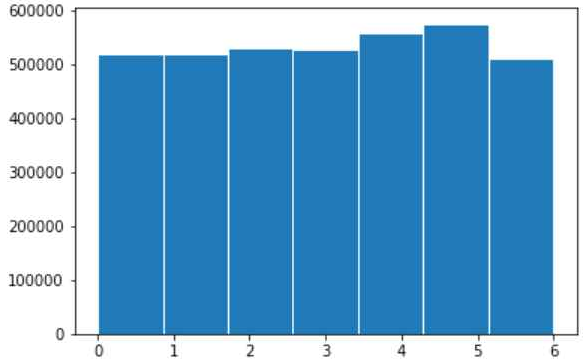
c. 연도별



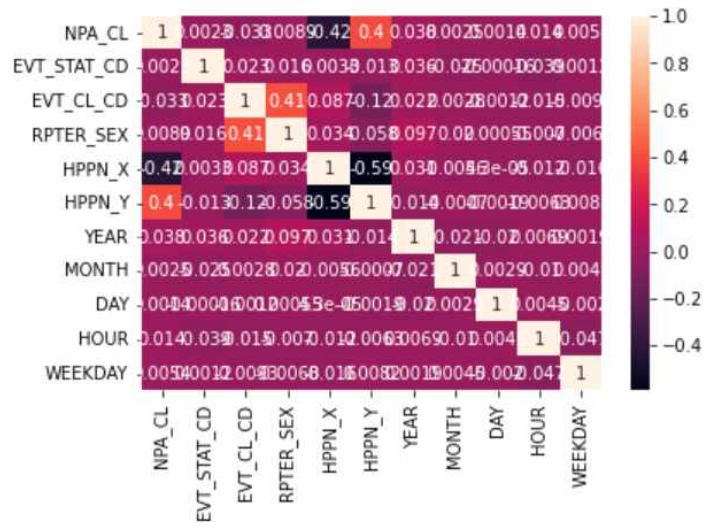
d. 성별별



e. 경찰청별



f. 요일별



g. 특성 간 상관관계

2-1-a. 시간별 분석

train 데이터를 'HOUR' 열을 기준으로 그래프를 그려본 결과, 새벽 5-6시 경에 가장 발생률이 적었고, 16-18시 경과 21-22시 경에 가장 높은 발생률을 보였다.

2-1-b. 월별 분석

train 데이터를 'MONTH' 열을 기준으로 그래프를 그려본 결과, 6-8월에 가장 높은 발생률을 보였고, 2월에 가장 낮은 발생률을 보였다. 추가적으로, 2023년의 경우에는 1월의 데이터만 존재하기 때문에 이를 고려하여 2023년을 제외한 데이터에도 같은 기준으로 적용해 보았을 때, 1월의 발생률은 5-9월과 비슷한 수치에서 12월과 비슷한 수치로 현저히 떨어진 것을 확인할 수 있었다.

2-1-c. 연도별 분석

train 데이터를 'YEAR' 열을 기준으로 그래프를 그려본 결과, 데이터가 상대적으로 부족한 2023년을 제외하면 거의 동등한 분포를 보였다.

2-1-d. 성별별 분석

train 데이터를 'RPTER_SEX' 열을 기준으로 그래프를 그려본 결과, 1(남자), 3(불상), 2(여자) 순으로 발생률이 높다는 것을 볼 수 있었다.

2-1-e. 경찰청별 분석

train 데이터를 'NPA_CL' 열을 기준으로 그래프를 그려본 결과, 19(충남), 13(대전), 31(세종) 순으로 발생률이 높은 것을 알 수 있었다.

2-1-f. 요일별 분석

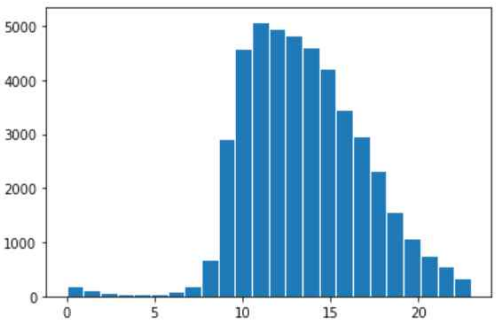
다른 요일들에 비해 4(금요일), 5(토요일)이 미세하게 많은 발생 건수를 갖는다. 하지만 유의미한 차이는 아닌 것으로 보이며, 주목할 만한 요일별 특징 역시 보이지 않는다.

2-1-g. 특성 간 상관관계 분석

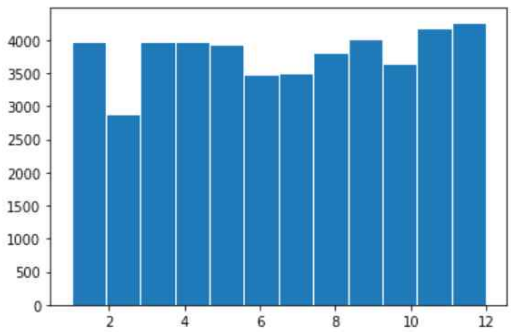
유의미한 상관관계를 가지는 특성 쌍이 존재하지 않는다.

2-2. 보이스피싱 자료 분석

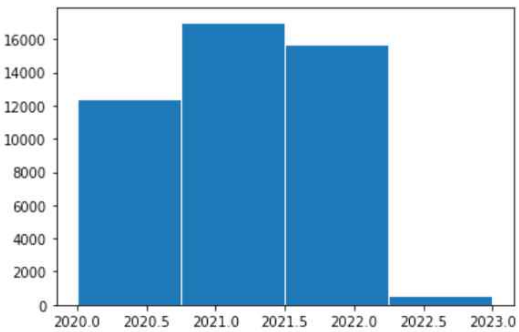
train 데이터에서 EVT_CL_CD이 215인 보이스피싱 데이터만 뽑아내 이를 train_215 데이터로 처리하였다.



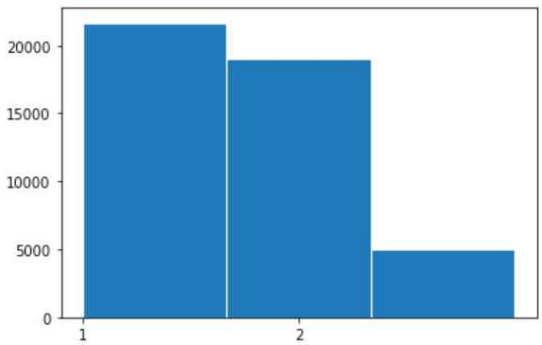
a. 시간별



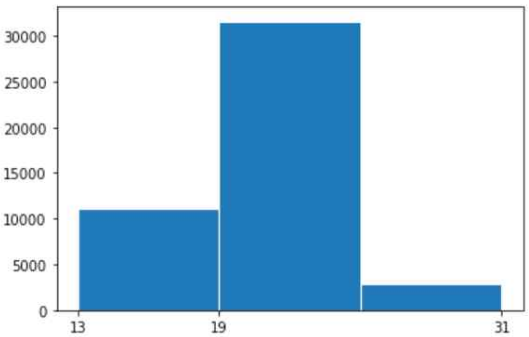
b. 월별



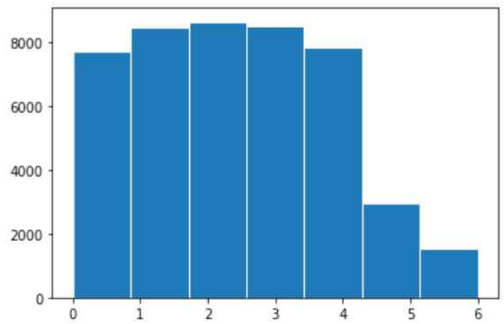
c. 연도별



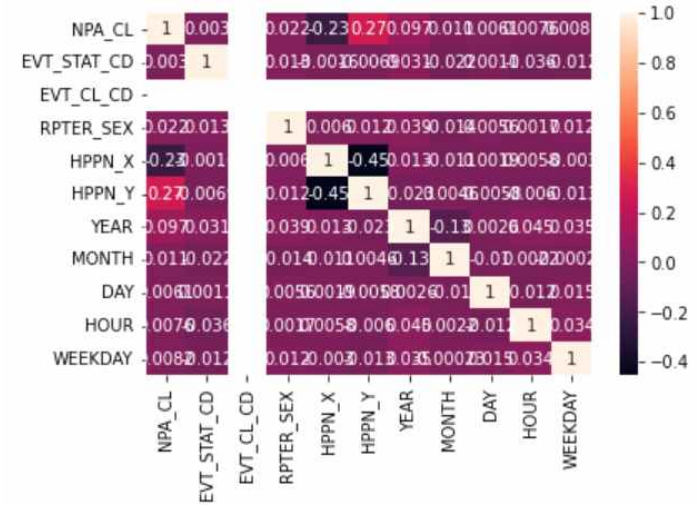
d. 성별별



e. 경찰청별



f. 요일별



g. 특성 간 상관관계

2-2-a. 시간별 분석

train_215 데이터를 'HOUR' 열을 기준으로 그래프를 그려본 결과, 새벽 3-4시 경에 가장 발생률이 적었고, 12-14시 경에 가장 높은 발생률을 보였다. 이와 더불어 전체 자료의 분석에 비해 밤 시간대의 발생률이 적어진 것과 전체적인 차이가 보다 극명해진 것 또한 확인할 수 있다.

2-2-b. 월별 분석

train_215 데이터를 'MONTH' 열을 기준으로 그래프를 그려본 결과, 2월에 가장 낮은 발생률을 보이고, 유의미한 경향을 보이지 않는다. 추가적으로, 이 데이터에도 2023년을 제외하여 같은 기준으로 적용해보았으나 큰 차이를 보이지 않았다.

2-2-c. 연도별 분석

train_215 데이터를 'YEAR' 열을 기준으로 그래프를 그려본 결과, 데이터가 상대적으로 부족한 2023년을 제외하면 2020년에 가장 낮은 발생률을 보였다.

2-2-d. 성별별 분석

train_215 데이터를 'RPTER_SEX' 열을 기준으로 그래프를 그려본 결과, 1(남자), 2(여자), 3(불상) 순으로 발생률이 높다는 것을 볼 수 있었다.

2-2-e. 경찰청별 분석

train_215 데이터를 'NPA_CL' 열을 기준으로 그래프를 그려본 결과, 19(충남), 13(대전), 31(세종) 순으로 발생률이 높은 것을 알 수 있었다.

2-2-f. 요일별 분석

0~4(평일)의 발생 횟수가 5~6(주말)보다 확연히 높은 것이 확인되었다. 이는 전체 범죄의 요일별 데이터(**2-1-f**)와 유의미하게 차이나는 보이스피싱 범죄만의 특징이다.

2-2-g. 특성 간 상관관계 분석

유의미한 상관관계를 가지는 특성 쌍이 존재하지 않는다.

3. 모델

전처리한 데이터를 바탕으로 회귀 모델을 작성하였다. 주어진 특성들 중, 예측시 이점이 있을 것으로 보이는 '사건종별코드(EVT_CL_CD)', '발생좌표(HPPN_X, HPPN_Y)' 특성에 대해 모델을 제작하였다.

3-1. 보이스피싱 여부 예측 모델

사건종별코드를 나타내는 특성인 'EVT_CL_CD'를, 다른 특성들로 예측하는 모델을 학습시켰다.

3-1-a. 모델 정보 요약

- **모델** : decision tree classifier
- **성능 지표** : 0.994 (f1-score)
- **hyperparameter** :

hyperparameter	value
criterion	gini
max_depth	2
min_samples_leaf	1

3-1-b. 학습 과정 설명

gaussian naive bayes classifier, decision tree classifier, multi-layer perception 세 가지 모델을 후보로 두고, 적절한 성능 지표를 채택하여 결정한다.

1) 성능 지표 채택

EVT_CL_CD가 215면 1, 나머지는 0으로 인코딩하였다. 1보다 0의 개수가 압도적으로 많은 데이터 불균형이 존재하였다. 때문에 단순 accuracy로는 모델의 성능을 판단할 수 없다. 이에 모델의 평가 지표로 f1 score을 채택하였다.[2)]

2) hyperparameter tuning & 모델 선택

grid search 방식을 이용하여 최적의 hyperparameter 값을 찾았다. 각 모델에 관해 두 단계의 튜닝 과정을 거쳤다. 넓은 범위에서 좁은 범위로 값을 좁혀가기 위함이다. 이 과정을 통해 성능 지표를 최대화한 후 비교하였다.

모델	f1-score
Gaussian Naive Bayes Classifier	0.9940
Decision Tree Classifier	0.9940
Multi-Layer Perceptron	< 0.0001

최종적으로 0.994의 우수한 지표를 갖는 decision tree classifier를 선택하였다. 아래는 세 가지 모델에 대한 상세 설명이다.

2) f1-score이란 Recall(재현율; 실제 True인 것 중 True로 예측한 비율)과 Precision(정밀도; True라고 분류한 것들 중에서 실제로 True인 것의 비율)의 조화평균을 말한다. 0에서 1 사이 값을 가지며, 1에 가까울수록 좋은 모델임을 나타낸다. 데이터가 불균형할 때도 활용할 수 있는 평가 지표이다.

○ Gaussian Naive Bayes Classifier

- 우수한 지표를 기록하였다.
- 그러나 특징 간의 복잡한 관계가 있는 문제에 적합하지 않으며 비선형 관계를 잡아낼 수 없다. 따라서 복합적인 요인이 존재하는 실제 범죄 상황에 알맞지 않을 수 있다. 이 때문에 최종 모델로 선택하지 않았다.

○ Decision Tree Classifier

- 우수한 지표를 기록하였다.
- 이해와 해석이 간단하여 의사결정 프로세스에 유용하다. 즉 실제 상황에서 모델의 결과를 보고 해석하기 용이하다는 장점이 있다.
- 범주형 및 숫자형 기능을 모두 처리할 수 있는 특징이 있다. 실제 상황에서 주어지는 데이터와 적합하다고 볼 수 있다.
- 트리가 깊고 가지가 많은 경우 훈련 데이터를 쉽게 과적합할 수 있다. Hyperparameter tuning을 통해 이를 해결하려고 노력했다.

○ Multi-Layer Perceptron

- 지표가 0에 가까운 것을 보아, 수렴에 실패하였다. 이는 noisy하고 복잡한 데이터에 대하여 성능이 저하되거나 수렴 속도가 느려지는 특징 때문으로 보인다.

3-2. 위도, 경도 예측 모델

보이스피싱 데이터에 대하여 위도(HPPN_Y) 및 경도(HPPN_X) 특성을, 다른 특성들로 예측하는 모델을 학습시켰다.

3-2-a. 모델 정보 요약

- **모델** : Gradient Boosting Regressor
- **성능 지표** : 0.0562 (경도 MSE), 0.0441 (위도 MSE)
- **hyperparameter** :

hyperparameter	HPPN_X	HPPN_Y
learning_rate	0.1	0.1
max_depth	2	3
min_samples_leaf	1	1
min_samples_split	10	2
n_estimators	50	50

3-1-b. 학습 과정 설명

decision tree regressor, multi-layer perceptron, gradient boosting regressor 세 가지 모델을 후보로 두고, 최고 성능 지표를 내는 모델을 선택한다. X, Y 좌표 예측는 독립적으로 진행한다.

1) 인코딩 및 손실함수 채택

- 위도, 경도 결측치를 평균치로 채워주었다.
- 'NPA_CL', 'EVT_STAT_CD', 'RPTER_SEX', 요일 데이터는 one-hot encoding을 진행하였다.
- 연, 월, 일, 사건 종별 코드를 drop한다.
- 손실 함수로는 평균 제곱 오차(MSE)를 채택하였다.

2) hyperparameter tuning

grid search 방식을 이용하여 최적의 hyperparameter 값을 찾았다. 각 모델에 관해 두 단계의 튜닝 과정을 거쳤다. 넓은 범위에서 좁은 범위로 값을 좁혀가기 위함이다. 이 과정을 통해 손실 함수를 최소화한 후 비교하였다.

모델	HPPN_X의 MSE	HPPN_Y의 MSE
decision tree regressor	0.0562	0.0441
multi-layer perceptron	35.0256	0.0919
gradient boosting regressor	0.0562	0.0441

아래는 세 가지 모델에 대한 설명이다.

○ **Gradient Boosting Regressor**

- 형상과 대상 변수 사이의 비선형 관계를 처리할 수 있어 광범위한 회귀 작업에 유용하다.
- 특이치에 강하므로 대부분의 데이터에서 크게 벗어나는 데이터를 처리할 수 있다. 그러나 특히 트리 수가 많거나 데이터 세트가 클 때 훈련 속도가 느릴 수 있다.
- 지나치게 복잡성이 높은 경향이 있으며, 이는 너무 복잡하고 해석하기 어려운 모델을 만들 수 있음을 의미한다. 이로 인해 성능이 저하되거나 수렴 속도가 느려질 수 있다.

○ **Decision Tree regressor**

- 3-1-b에서 기술한 decision tree classifier과 같은 방식이다.

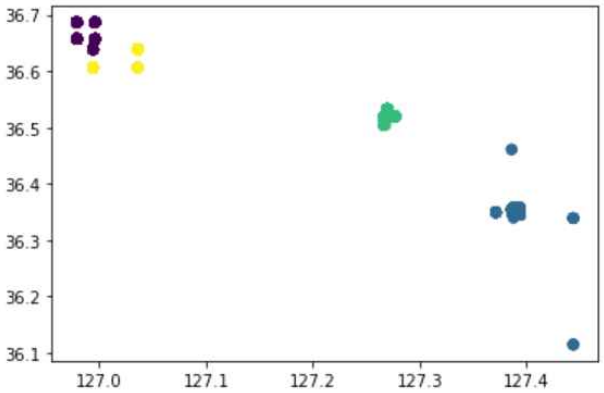
○ **Multi-Layer Perceptron**

- 3-1-b에서 기술한 multi-layer perceptron과 동일하다.

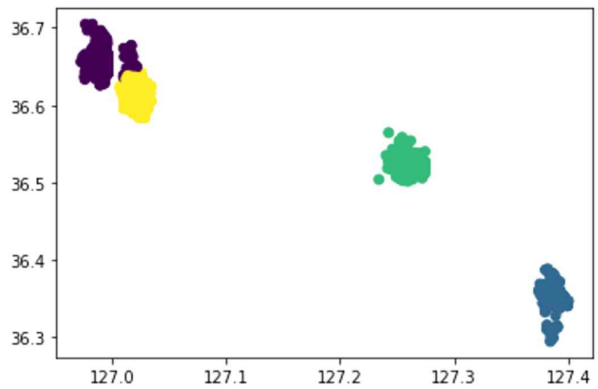
3-2-c. 군집화 & 모델 선택

예측 모델의 활용 가능성을 높이기 위해 경향성을 파악하고자 예측 지점들의 군집화를 knn 방식으로 진행한 뒤 시각화했다.(가로축:경도, 세로축:위도) 3-1-b에서 손실값이 일치했던 decision tree regressor, gradient boosting regressor 두 모델을 비교하여 최종 모델을 선택하고자 했다.

군집 수는 1부터 10까지의 값에 따라 inertia[3]를 계산한 뒤, elbow 이론[4]에 따라 4개로 지정하였다.



decision tree regressor, 군집 수=4

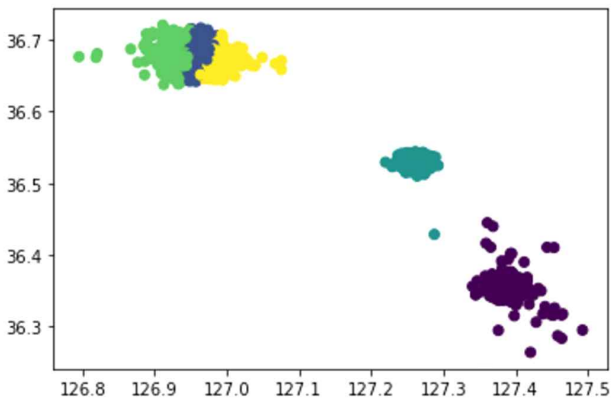


gradient boosting regressor, 군집 수=4

decision tree regressor의 경우, 특정 지점에 예측 데이터가 몰려 있어 적절치 않은 예측 모델임을 확인했다. gredient boosting regressor의 경우도 비슷한 양상을 보이긴 하나, 상대적으로 퍼져 있는 좌표로 예측하기 때문에 이를 더 나은 예측 모델이라고 결론지었다.

- mean imputation(결측치를 평균으로 채우는 과정) 생략 시

같은 지점에 예측치가 몰리는 현상을 해결하기 위해, 좌표 결측치가 존재했던 데이터를 전부 삭제하고 다시 gradient boosting regressor 모델로 발생 좌표를 예측하도록 했다. 이후 군집화를 다시 진행한 사진은 아래와 같다. 군집 수는 5개로 설정하였다.



3) inertia = 각 군집의 중심에서 군집에 할당된 데이터 포인트간의 거리를 합산한 것. 작을수록 촘촘한 군집임을 나타냄
4) inertia, 군집 수로 그래프를 작성한 뒤, inertia의 감소량이 꺾이는 지점의 군집 수를 채택하는 방법

이전보다 더 고른 분포를 보이는 것을 확인할 수 있었다.

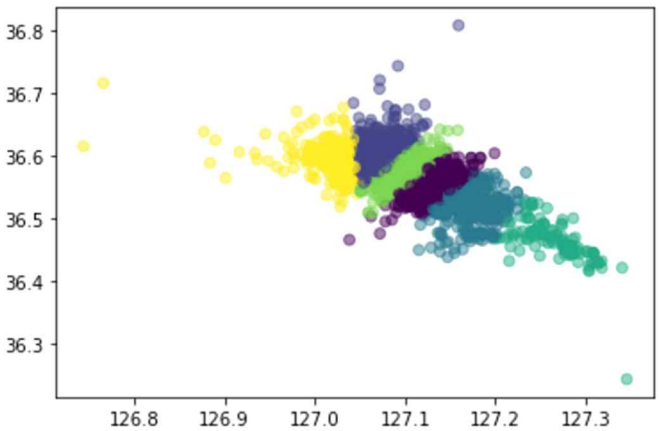
- NPA_CL 특성(경찰청 번호) 제거 시

지금까지의 군집화 경향을 봤을 때, 군집을 4개에서 5개로 설정하지만 크게 봤을 때 3가지로 나누어지는 것으로 보인다. 실제 세종, 대전 위치에 데이터 군집이 위치하는 것으로 보아, 데이터의 NPA_CL 특성이 발생 위치를 예측하는 데에 크게 관여하는 중이라고 해석할 수 있다. 이에 데이터에서 NPA_CL 특성을 전부 제거하고 예측을 진행했다.

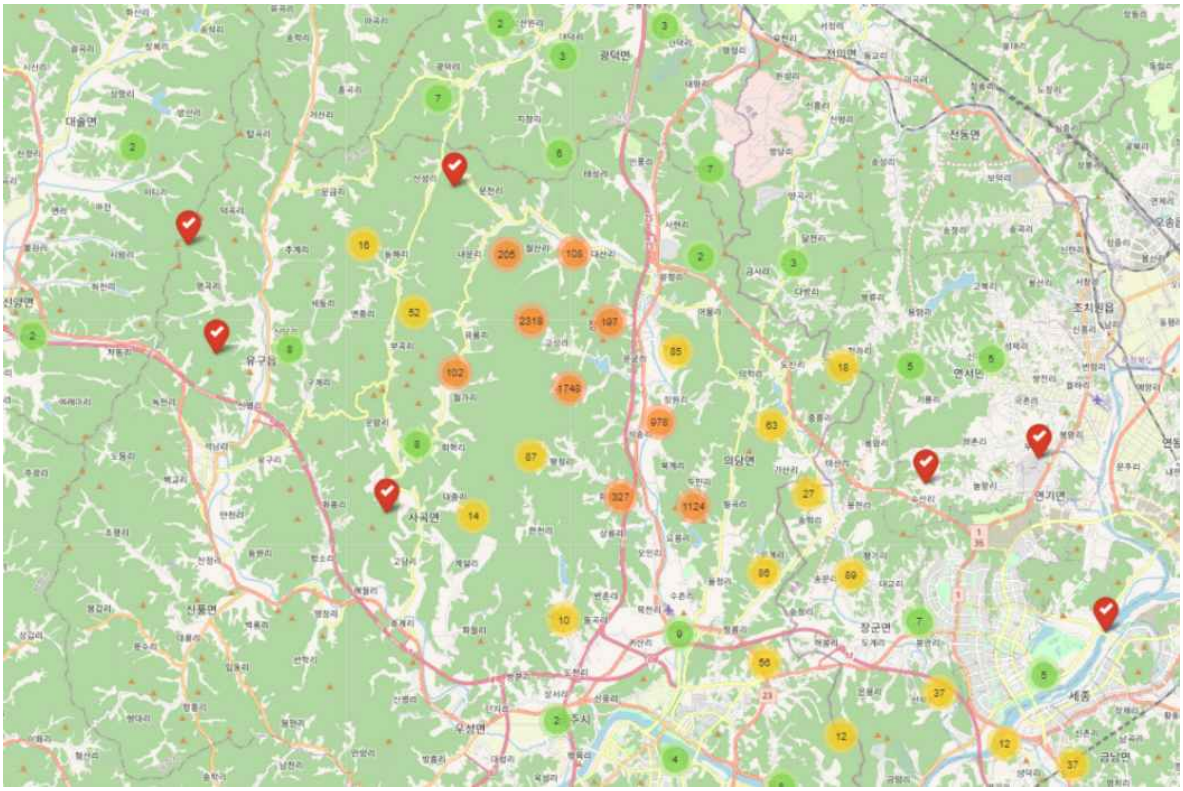
추가로, 튜닝 시간을 단축시키고 일반화 성능을 개선하기 위해 grid search에서 random search 방법으로 변경하였다. 해당 hyperparameter, MSE 및 군집화 결과이다.

hyperparameter	HPPN_X	HPPN_Y
learning_rate	0.05	0.046
max_depth	6	6
min_samples_leaf	5	4
min_samples_split	8	10
n_estimators	50	52

HPPN_X의 MSE : 0.0971, HPPN_Y의 MSE : 0.0674으로, 손실은 소폭 증가하였다,
그러나 군집화시 세 지점으로 밀집되는 현상이 크게 해결되었다.

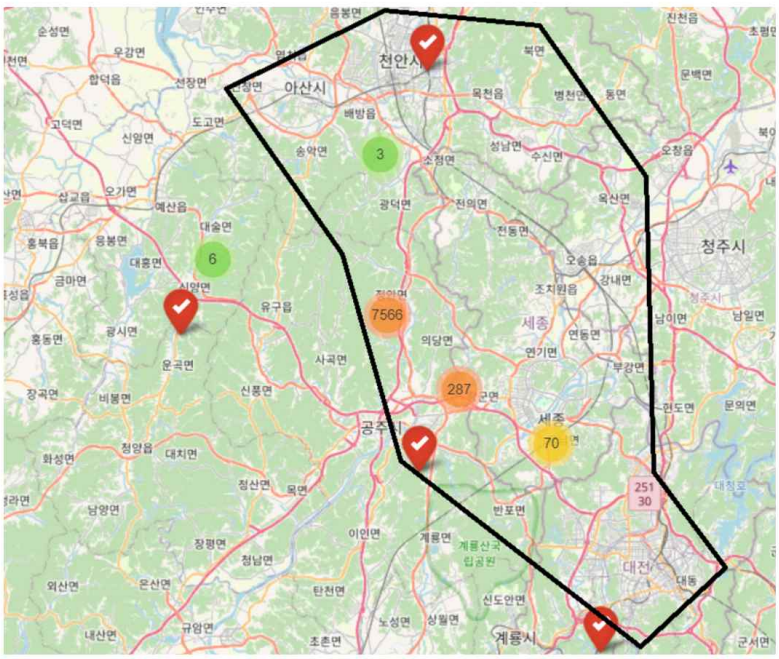


위 산점도의 해석이 용이하도록, 실제 지도 위에 나타내었다. python의 folium library 사용하였다. 공주시의 좌표를 중심으로 예산, 세종, 대전 등에 훨씬 더 넓고 고르게 분포하는 것을 볼 수 있다.



이를 실제 데이터와 비교하면 여전히 사건 수가 많은 천안, 아산 등에 예측값이 분포하지 않는 등 문제가 존재한다. 이는 데이터셋 자체의 한계인 것으로 파악된다.

군집화 결과를 바탕으로 대략적인 보이스피싱 취약 구역을 지도에 표시하면 다음과 같다.[5)]



5) 충남 전체 지도로 확대하니 가운데에 데이터 대부분이 몰려있는 것처럼 표시되었으나, 해당 구역 확대 시 고르게 분포되어 있다.

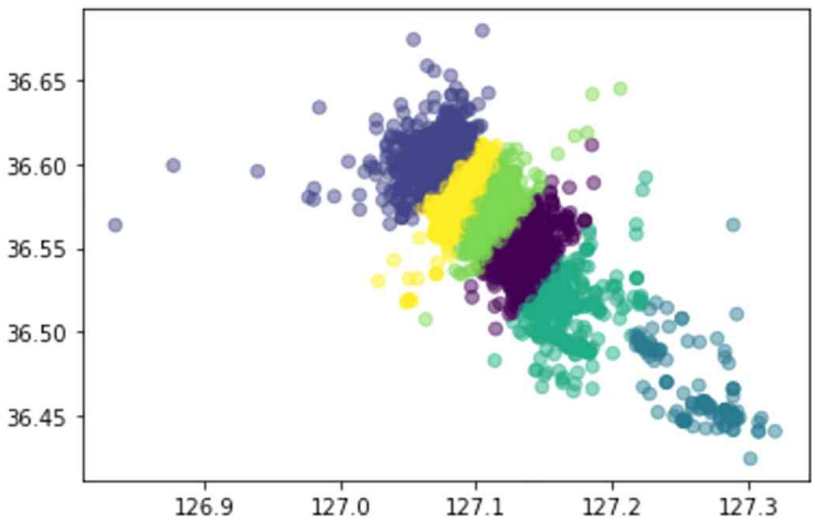
해당 구역 결정 과정 :

- 위도와 경도의 MSE를 실제 거리로 환산하면 가로 약 34.6km, 세로 약 23.2km의 오차라는 의미이다.
- 공주시 정안면으로부터 왼쪽은 인구가 적고 오른쪽은 인구가 많기 때문에, 정안면을 중심으로 오른쪽으로 치우친 도형을 그렸다.
- 도형의 좌우 폭은 약 34.6km에 미치지 않지만, 공주시 정안면에서 오른쪽으로 약 34.6km를 이동한 경우 충청북도가 다수 포함되고 연구 주제와 맞지 않는 지역이기 때문에 천안, 세종, 대전의 경계 부분까지만 확장하였다.
- 공주시 정안면으로부터 위아래 약 23.2km 차이나는 지점을 도형의 경계로 결정하였다.

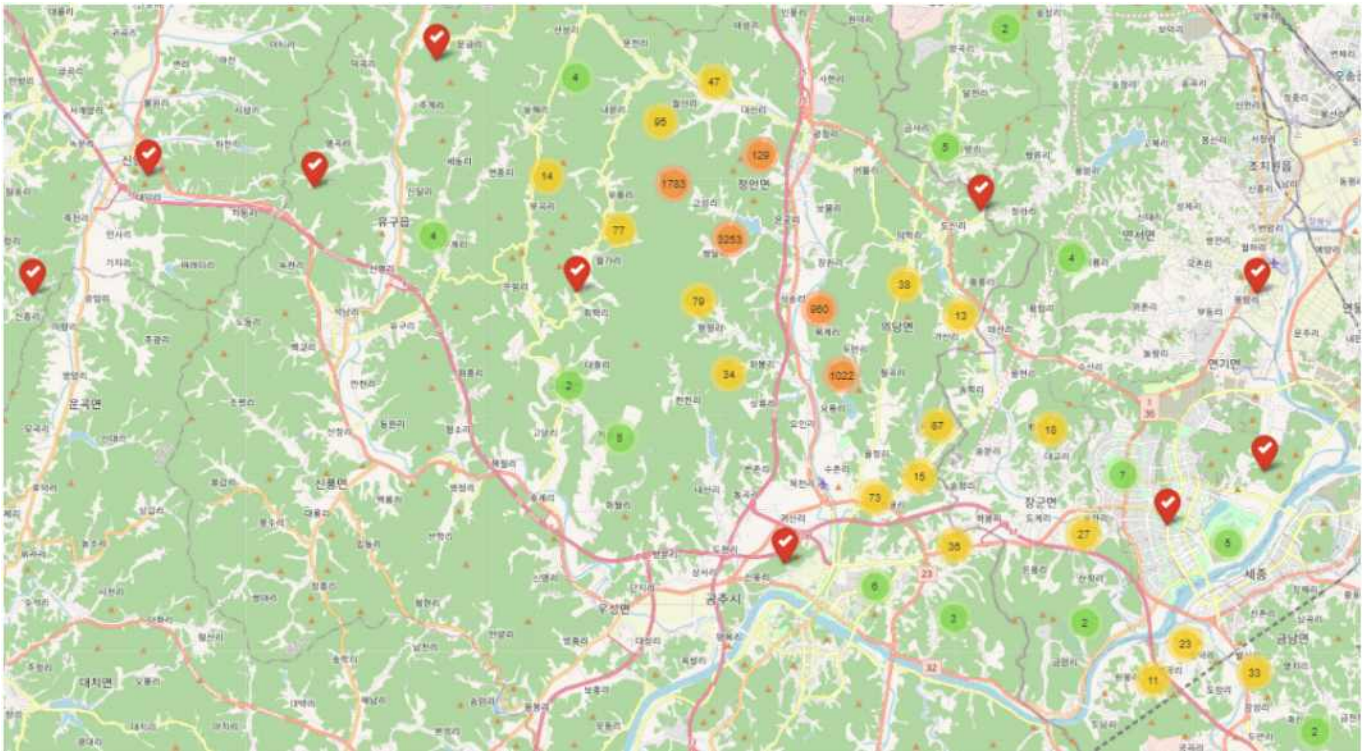
- 데이터 정규화(normalization)에 관하여

모델의 예측 목표값인 위도와 경도를 제외한 나머지 특성들은 모두 이산형이기 때문에, 정규화를 하지 않아도 학습에 지장이 없다고 판단했다. 때문에 이전까지의 실험은 주어진 데이터셋의 그 자체 값으로 진행하였다. 하지만 일반적으로 정규화를 진행하기 때문에, 본 연구에서도 진행한 결과를 첨부하겠다.

바로 전 실험과 똑같은 과정을 거치되 training set의 위도, 경도를 정규화한 후 모델을 학습시킨 뒤, test set에 대하여 해당 모델을 적용한다. 그 결과값을 다시 위도 경도로 변환하기 위해 다시 scikit-learn의 inverse_transform 함수를 이용해 복구한 뒤 시각화한다.



조금 더 고르게 분포된 것으로 보이나, 아래 지도에 시각화한 결과를 보면 개선되지 않았음을 확인할 수 있다.



○ 결과 해석 및 시사점

1. 데이터

유의미한 feature로 보이는 '발생 시간', '경찰청별' 항목에 대한 해석이다.

1-1. 범죄 발생 시간별 (2-1-a, 2-2-a에 해당함)

- 0시부터 07시까지의 보이스피싱 범죄 건수가 전체 발생 건수에 비해, 또 보이스피싱의 다른 시간대에 비해 현저히 낮다. 실상 거의 존재하지 않는 수준인데, 이는 가해자와 피해자가 전화기로 연결되어야 하는 보이스피싱 범죄의 특성 때문인 것으로 보인다.
- 보이스피싱 범죄 발생 건수에서, 09시경을 기점으로 급격히 발생 건수가 증가하고 15시부터 24시까지 점차 감소하는 추세를 보인다.

1-2. 경찰청별 (2-2-e에 해당함)

대전, 충남, 세종의 보이스피싱 신고 건수는 약 3년간 11176, 31595, 2936건이다.
이를 연평균으로 환산한 뒤 해당 지역의 1인당 연간 보이스피싱 신고 수를 계산하여 비교 지표로 활용한다.[6] 계산 결과는 다음과 같다.

6) 즉, '발생 건수 ÷ 3(년) ÷ 인구수' 식으로 지표화했다. 소수점 4자리 이하는 버림한다. 인구수는 2022년을 기준으로 하여 대전, 충남, 세종 순서대로 약 153만, 206만, 38만 명이다.

지역	**비교 지표** (연간 보이스피싱 발생 건수)
대전	0.0024
충남	0.0051
세종	0.0025

대전, 세종, 충남의 비율이 약 1:1:2 정도이다. 충남 지역의 범죄 건수가 상대적으로 높은 것이 주목할만한 점이다. 충남 지역의 보이스피싱 담당 경찰력을 강화하고, 시민 대상의 범죄 예방 교육, 공익광고 등을 시행하는 등 접근성 높은 방안을 강구한다면 유의미한 효과를 얻을 수 있을 것으로 보인다.

1-3. 요일별 (2-1-f, 2-2-f에 해당함)

모든 범죄 발생 데이터에서는 요일별 구분이 큰 의미가 없는 반면, 보이스피싱 범죄는 주말에 큰 폭으로 발생 건수가 적었다. 평일과 주말의 발생 건수는 대략적으로 16 : 3의 비율로, 이 수치를 요일별 경찰력 분배에 활용 가능할 것으로 보인다.

2. 모델

2-1. 보이스피싱 범죄 여부 예측 모델

피해자가 자신이 처한 범죄 상황을 인지하지 못하는 경우에, 우수한 성능을 가진 모델이 범죄 상황을 정확히 파악하고 대처할 수 있도록 도울 수 있다.

2-2. 위도 경도 예측 모델

주어진 실제 데이터셋에서도 위도, 경도 결측치가 다수 존재했다. 신고자가 GPS 기능을 활성화할 수 없거나 활성화 방법을 알지 못하는 경우. 신고자의 위치 전달이 어려운 상황 등에 신고자의 위치를 대략적으로 특정할 수 있는 방법이 될 것이다.

○ 기대효과

‘결과 해석 및 시사점’의 연장선상에 있는 내용이므로, 글번호를 유지하여 작성하였다.

1. 데이터(분석을 통해 얻을 수 있는 기대효과)

결과 해석에서 진행한 경향성 파악을 통해 경찰력을 효율적으로 분배할 수 있다. 특히 요일별 분석의 경우 평일과 주말 간에 유의미한 범죄율 차이가 있으므로 수사망 확대, 인력 재배치 등으로 효과를 볼 수 있을 것이다.

2. 모델

2-1. 보이스피싱 범죄 여부 예측

- 실제 신고시 어떠한 특성이 누락된 경우에, 즉 모든 정보를 수집하기 어려운 상황에 모델을 적용하여 해당 특성을 예측할 수 있다. 특히 범죄 유형을 특정하기 어려울 때 높은 정확도로 그 유형을 예측하는 모델을 적극적으로 이용할 수 있을 것으로 보인다.
- 신고자의 데이터 수집이 즉시 이루어진다면, 담당 부서를 신속하게 파악하고 사건을 인도할 수 있기에 처리 및 대응 속도가 올라갈 것으로 보인다.
- 경찰 측에서 수집, 제공하는 공공 데이터의 종류가 다양해지고 질 높아진다면 경찰력에 더욱 힘을 실어줄 수 있는 분석이 가능할 것으로 보인다. 개척 여지가 많은 지점이다.

2-2. 위도/경도 예측 및 군집화

- 취약 구역을 그려봤을 때, 공주시 주변으로 대전-세종-천안-아산 지역에 범죄 예측 지점이 많이 분포되어 있다. 따라서 공주시 중심으로 보이스피싱 경찰력을 증원하는 조치가 필요할 것으로 보인다. 현재 각 경찰청 별로 집행되고 있는 부서별 인력을 알 수 없기에, 본 보고서에서는 구체적인 증원/감원 수치를 제시하지 못할 것으로 보인다.
- 신고자의 위치 정보가 누락된 경우에, 다른 정보들을 입력으로 하여 대략적인 담당 구역을 특정하는 용도로 활용 가능하다.

Ⅳ. 기타

○ 건의 사항

- 데이터 형태의 구조적인 문제점

데이터가 신고 직후에만 수집되는 것들로만 이루어져 있다. 이러한 데이터는 분석이 한정적일 수밖에 없고, 선제 대응이라는 취지에 부합하는 모델을 구현하는 것에 한계가 있을 것으로 보인다. 활용할 데이터를 다양화하고 질을 높인다면 더욱 고도화된 모델과 분석이 가능할 것이다.

○ 활용 데이터 및 참고 문헌 출처 (필수)

참고 문헌

조준택, 김강일, 박현호, 범죄예방을 위한 112신고 자료의 활용방안 연구(한국치안행정논집 제15권 제1호), p6, 2018

최재훈. (2018), 기계학습을 활용한 데이터 기반 경찰신고건수 예측, 한국빅데이터학회지 제3권 제2호, pp. 101-112.