10월 5일

- 정답을 맞추지 못한 로그들을 분석
- 새로운 공격 유형(web shell) 발견
- 현재 공격 분류를 어떤 기준으로 하고 있는지 궁금해짐
    - CatBoost Plot tree를 이용해서 분류 기준 확인 필요
    - https://catboost.ai/en/docs/concepts/python-reference_catboost_plot_tree
- 공격 유형(recon, exec, post) 에 해당하는 공격이 포함되면 가중치를 부여하는 방식을 적용하면 어떨지?
- catboost 모델의 eval_metric을 "TotalF1"이 아닌 "MultiClass"로 설정하면 모델링 시간(6분->25분)이 약 4배정도 걸리지만 macro-precision이 0.9809로 약간의 성능 상승이 이루어졌다. (s_ip, d_ip, host ip, ip2*3*4, payload 제거 후 모델링 수행)
- cat.pkl(10/5 18:00 쓸모없는 ip 및 변수 제거 후 새로운 공격유형이 들어간 모델)
- 공격분류에 따른 라벨 변경은 모델이 예측한 값에 비해 정밀도가 떨어졌다.

10월 6일

- 8개의 공격분류변수와 8개의 공격분류변수값을 붙인 attack_pattern 변수를 사용한 모델이 macro_precision 기준 0.9809로 가장 높은 성능을 가졌다.
- cat.pkl(10/6 14:00 8가지 공격분류 + attack_pattern)
- 성능 향상을 위해 IP와 Payload에서 피쳐를 뽑고 적용하는 것에 집중했었음. 현재는 risk는 별도로 고려하지 않았음. 혹시나 싶어 risk와 공격 분류(1~3)과 연관이 있을지가 궁금해졌고 상관 계수가 0.5로 나옴

10월 7일

- 회의해야 할 내용
    1. 페이로드를 이용하여 분류된 공격이 정확한지?
       - 현재는 패턴 매칭으로 하고 있어서 문맥을 고려하지 않아 정확하지 않은 문제가 있음. 얼마나 정확하지 않은지는 50,000개 모두를 대상으로 직접 확인해봐야 알 수있는 문제 (이건 아직 보류)
    2. 페이로드가 같은 상황에서 공격 유형(class)이 다른 경우는 어떻게 처리할지?
       - 페이로드가 같으면 ip, risk는 다름. 이 정보들로 판단을 해야하는데 어떤 피쳐가 영향을 주는지 상관 관계 파악 필요

- 1) ip의 특정 클래스(또는 대역)가 공격 유형 판정에 영향을 주는지?
- 2) risk가 공격 유형 판정에 영향을 주는지?
- ip 값 자체는 다르겠지만 클래스가 같고, port도 같고, risk도 같다면 어떤 기준으로 공격 유형을 판단할지? 또는 버릴지?
- 중복되는 페이로드 중에서 오탐으로 표시된 비율이 약 33%이므로 중복된 페이로드를 처리할 방법 고민 필요

10월 8일

- 페이로드가 같은 상황에서 공격 유형(class)이 다른 경우는 어떻게 처리할지?
  - 1) ip의 특정 클래스(또는 대역)가 공격 유형 판정에 영향을 주는지?
  - 2) risk가 공격 유형 판정에 영향을 주는지? **상관 관계 없음**

- catboost model1이 예측한 값을 변수로 사용하여 catboost model2를 모델링하는 아이디어를 생각해보았다. 98%의 잘 예측한 라벨과 2%의 잘못 예측한 라벨을 모델이 학습하여 성능이 좋아지지 않을까?(동일한 성능)

10월 9일

- 중복되는 페이로드 처리를 위한 방법으로 IP에서 추가적인 피쳐를 추출할려고 생각함
- IP의 위치 주소를 파악하기 위한 방법을 찾다가 한국인터넷정보센터에서 해외 IP 현황 (https://xn--3e0bx5euxnjje69i70af08bea817g.xn--3e0b707e/jsp/statboard/IPAS/ovrse/natal/IPaddrBandCurrent.jsp) 을 발견
- 가나에 할당된 IP 대역만 봐도 한 대역 범위가 넓은게 아니라 짧은 범위로 여러개 갖고 있는 것을 확인. 그렇다면 IP 클래스(A~D)만으로는 상관 관계가 애매하게 나올 수 있다고 생각이 듬(비례 관계가 아니므로)
- 전세계에 알려진 IP 갯수가 23만개 였고 IP 쌍(출발지, 목적지)을 식별하는데 3초 정도 걸림
- 전체 로그 갯수가 5만 이므로 3초 x 5만 = 15만초 = 1일 17시간 40분 ~~(COC 인줄?)~~
- 그래서 다른 방법 찾던 중 IP의 위치를 조회해주는 API 발견
- 이전에 진행해둔 위치 데이터 csv 파일에서 API로 조회한 값과 다른 것을 발견. 그럼 한국인터넷정보센터는 신뢰성이 좀 떨어진다는 얘기

10월 10일

- 무료 API 찾다가 아주 좋은 사이트(https://www.abstractapi.com/what-is-my-ip-address-and-location) 발견
- 여기는 가입해도 이메일 검증할 필요 없이 API 키를 얻을 수 있고 한 API 키당 3만개 제한 + 1초마다 요청 제한있는 것으로 보임
- 그러면 API 키는 무제한 생성할 수 있고, 시간 측정 해보니 100개에 4분 15초 걸림
- 병행으로 진행하면 좋은데 API 기준으로 1초 제한 재는게 아니라 IP 기준으로 시간 제한 재는듯?