Wydział Matematyki i
Nauk Infomacyjnych

Android Malware Detection

Michał Binda i Mikołaj Mróz

Wstęp

Eksploracja Danych

Preprocessing

Implementacja Modelu

Podsumowanie

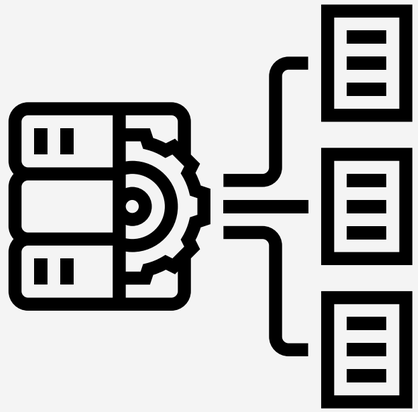
Naszym model miał przewidywać, do której kategorii z 4 należy Androidowe oprogramowanie na podstawie wielu innych danych tj. Source IP, Destination IP itd.

Kategorie były następujące:

- Android_Adware - Adware jest formą złośliwego oprogramowania, które ukrywa się w urządzeniu i wyświetla reklamy
- Android_SMS_Malware - SMS Malware to każde złośliwe oprogramowanie dostarczane ofiarom za pośrednictwem wiadomości tekstowych
- Android_Scareware - próbuje przestraszyć użytkowników, aby myśleli, że ich urządzenie zostało zainfekowane wirusem, a następnie zachęca ich do szybkiego pobrania programu, który to naprawi
- Benign - nieszkodliwe oprogramowanie

Preprocessing

- Usunięcie niepotrzebnych spacji na początku nazw kolumn
- Zastąpienie braków danych modą z danej kolumny
- Usunięcie kolumn, które mają stałą wartość
- Naprawa niepoprawnych adresów IP
- Zmodyfikowanie IP (podzielenie go na 4 oddzielne kolumny)



Problem, który dotyka miliardy

Szkodliwe oprogramowanie może zaszkodzić każdemu z nas.
Jak sobie z nim radzić?

Wydział Matematyki i
Nauk Infomacyjnych

Który Model Wybrać?

Wydział Matematyki i Nauk Infomacyjnych

K - Nearest Neighbor

- Jest to algorytm klasyfikacji lub regresji, który przyporządkowuje dane do klas na podstawie ich sąsiedztwa.
- Działa na zasadzie znajdowania najbliższych sąsiadów dla nowych danych i przypisywania do nich klasy na podstawie większościowego głosowania.

Decision Tree

- Jest to model predykcyjny, który reprezentuje decyzje lub zasady decyzyjne w formie drzewa.
- Na podstawie cech wejściowych przyporządkowuje dane do odpowiednich klas lub podejmuje decyzje na podstawie warunków logicznych.

Random Forest

- Jest to zbiór drzew decyzyjnych, który działa na zasadzie łączenia wyników wielu drzew w celu uzyskania bardziej stabilnych i dokładniejszych predykcji.
- Działa na zasadzie losowego wyboru podzbiorów danych i budowy drzew na ich podstawie, a następnie łączenia wyników drzew w celu uzyskania końcowej predykcji.

XGB

- XGBClassifier to algorytm uczenia maszynowego oparty na drzewach decyzyjnych, który jest szczególnie skuteczny w zadaniach klasyfikacji binarnej i wieloklasowej.
- Wykorzystuje on metodę gradientowego zwiększania drzew (Gradient Boosting),

Porównanie wyników

$$Precision = \frac{TruePositive}{(TruePositve + FalsePositive)}$$

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

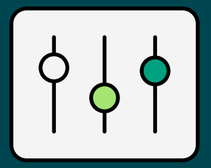
$$F - Measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

Porównanie wyników

Metryka	Random Forest	Decision Tree	XGB
Accuracy	0.68	0.67	0.74
Precision	0.67	0.62	0.75
recall	0.67	0.62	0.74
F - measure	0.67	0.62	0.74

Testy dla zbioru testowego

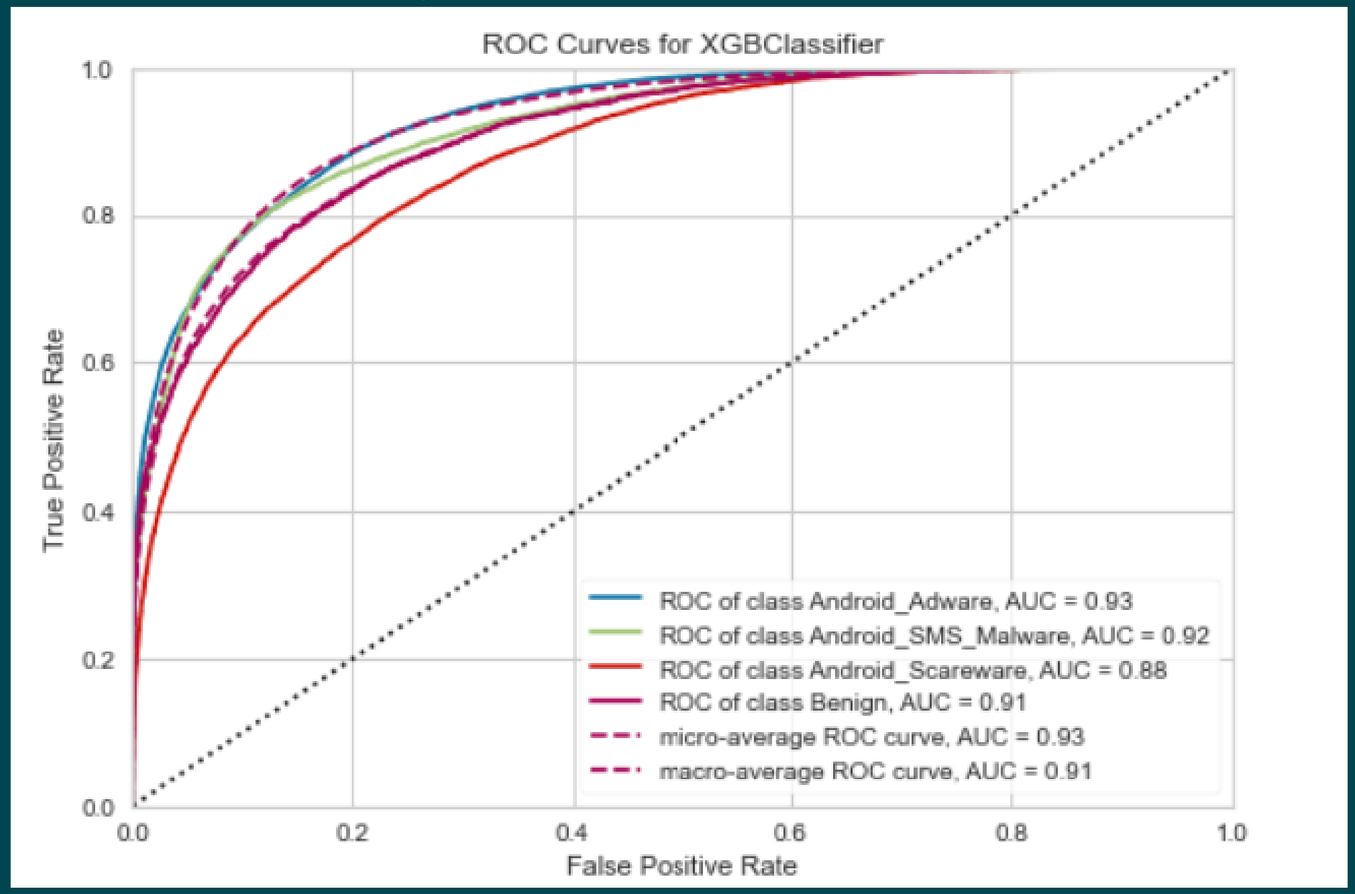
Wybór XGBoost



XGBoost jest jednym z najlepszych algorytmów uczenia maszynowego, ze względu na wiele zaawansowanych funkcji, co pozwala na uzyskanie wysokiej skuteczności w predykcji, szybkie przetwarzanie dużych zbiorów danych, integrację z innymi narzędziami i platformami. W przypadku identyfikacji złośliwego oprogramowania na platformie Android, XGBoost może przynieść doskonałe rezultaty, ze względu na dużą stabilność i niezawodność, a także na możliwość obsługi wielu rodzajów danych wejściowych.

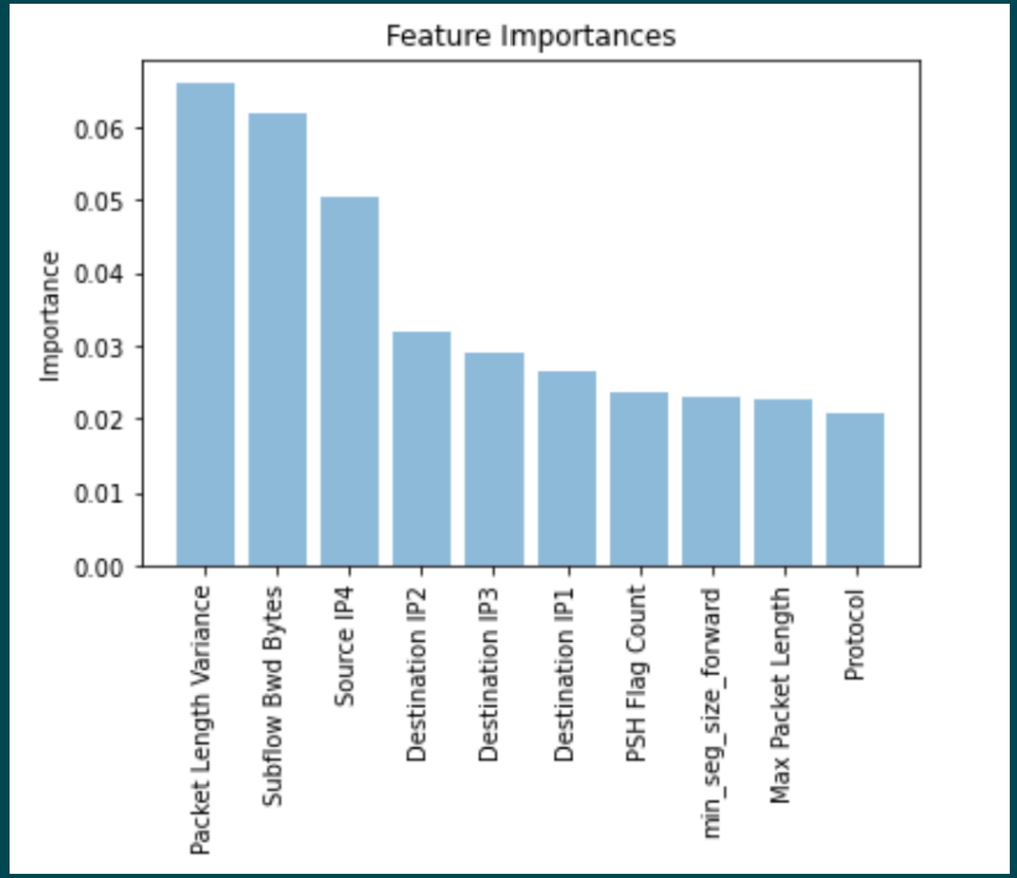
Wydział Matematyki i
Nauk Infomacyjnych

Porównanie wyników



Testy dla zbioru testowego

Najważniejsze cechy



Testy dla zbioru testowego

Interpretacja modelu

W przypadku modelu uczenia maszynowego, który ma za zadanie klasyfikować oprogramowanie jako złośliwe lub niezłośliwe, zmienne takie jak Packet Length Variance, Source IP i Destination IP mogą być bardzo ważne i predykcyjne.

Packet Length Variance to miara różnorodności długości pakietów sieciowych w transmisji. Złośliwe oprogramowanie często wykorzystuje specjalnie spreparowane pakiety, aby ukryć swoją aktywność lub wykonać atak, co może prowadzić do większej wariancji długości pakietów.

Source IP i Destination IP to adresy źródłowe i docelowe w transmisji sieciowej. Złośliwe oprogramowanie często łączy się z komputerami, które są znane z hostowania innych złośliwych działań lub z botnetów. Wykrycie takiego połączenia może wskazywać na obecność złośliwego oprogramowania.

Wszystkie te zmienne są związane z cechami sieciowymi transmisji, które mogą być wykorzystywane przez złośliwe oprogramowanie w celu ukrycia swojej aktywności lub wykonywania ataków. Dlatego też, ich uwzględnienie w modelu uczenia maszynowego może być bardzo pomocne w poprawieniu jego skuteczności w wykrywaniu złośliwego oprogramowania.

Wydział Matematyki i
Nauk Infomacyjnych