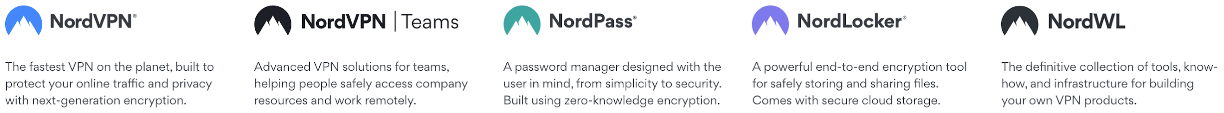
NORD
SECURITY

Company

We are leaders in the field of online privacy and security. Over 15 million users trust us to keep their data safe on the internet.

Nord Security was born as a passion project, and our drive is reflected in our work. Basically, we want one thing only — to give true online privacy and security to as many people as we can. At the moment, we have built 5 different cybersecurity products, which are used by millions of people worldwide.



NordVPN®

The fastest VPN on the planet, built to protect your online traffic and privacy with next-generation encryption.

NordVPN | Teams

Advanced VPN solutions for teams, helping people safely access company resources and work remotely.

NordPass®

A password manager designed with the user in mind, from simplicity to security. Built using zero-knowledge encryption.

NordLocker®

A powerful end-to-end encryption tool for safely storing and sharing files. Comes with secure cloud storage.

NordWL

The definitive collection of tools, know-how, and infrastructure for building your own VPN products.

Cybersecurity is all about hard-work, a modern technology stack, speed, a constant desire to learn, and above all, vigilance in keeping every last asset safe and sound.
Over less than a decade, we have grown from 4 friends to hundreds of cybersec experts — and we don't plan on stopping.

WE BELIEVE in:

- **Serving people through technology.** We build tools to make everyone feel safe and free on the internet.
- **Treasuring trust.** Respect for our users' data is ingrained in every process at Nord Security.
- **Focusing on what's real.** We aim to make security accessible to people who don't really do technology.
- **Putting privacy first.** We design privacy and security solutions with zero-knowledge infrastructure, so your details are safe with us.

Task

Each day we have more and more malware, which is affecting people across the world more often than ever. The times when malware was just for fun is over. Cyber criminals are utilizing the malware to do concentrated attacks on institutions, to collect the data about infected victims, steal the banking accounts, or encrypt the victims data and ask for the ransom.

Old fashioned antivirus companies with huge malware analysis departments can’t keep the pace to create the antidotes for all the malware. Everyone's eyes are now on new technologies and how to make the antimalware solutions more robust and fast.

This task is for the data scientists that are brave enough to dive into computer science.

Your job is to empower machine learning by separating the files into clean (benign) or malicious. While working on this task, you will investigate and learn how to preprocess data for the model. How much data is needed to increase the accuracy? How to reduce the number of false positives?

The goal is to have a file preprocessing pipeline with a working model, predicting the malware with the highest accuracy on a given dataset. And, of course, having the minimum number of false positives at the same time.

Goals

Main goals:
* Build a NN model separating malware files from benign files. For that, employ the provided Windows PE files dataset by extracting features from them (e.g. using the pefile python library). Baseline model should show results better than the random model. (30%)
* Set model performance metrics. Split the dataset to train and test datasets, and explain model performance using selected metrics. Prepare a confusion matrix and describe it. (30%)
* By detecting malware on a user's computer, we have to be of some kind of certainty that we do not label clean files as "malware". That is an unwanted behavior. This can be achieved by decreasing the value of False Positive Rate (FPR). In which ways can it be done? What issues does it create and how to deal with them? Code it. (20%)
* General understanding of data science topics behind the project (10%)
* General understanding of the field (10%)

Bonus goals:
* Text is valuable information in binary files, which can be used as features to detect malicious files. Try to extract it from raw files and use it as features. How does it improve model performance and why? (10%)
* Windows PE files basically consist of extracted metadata and binary data. Are there ways of using the binary part? Can the binary part be translated to some more meaningful representation? What are techniques of binary interpretation in a field of malware detection? Describe them. (10%) Try to use any of the described binary interpretations to improve classification results. (20%)

Note: solution should be presented in a .ipynb notebook containing both the solution/code and detailed reasoning/explanation behind your thinking.

Dataset

Dataset can be found here:
http://s3-nord-challenge-data.s3-website.eu-central-1.amazonaws.com/
Catalog 0 - clean files
Catalog 1 - **Malware - real malware, be careful!**