DOMAINTOOLS

# Technical Brief: Domain Risk Score

Proactively uncover threats using DNS and data science


## Introducing Domain Risk Score

Threat actors worldwide register and weaponize domains every day as part of phishing, malware, and spam campaigns. The goal of network defense is to identify and flag these domains registered with "malicious intent" before they are weaponized. Such advance detection and categorization has many applications, including perimeter or host defenses, on-network threat discovery, or forensic review. Our Domain Risk Score can be used to augment an organization's existing threat intelligence processes. One can think of domains with a high risk score as belonging on a "domain watchlist"—they are domains which the algorithms indicate may become dangerous in the near future.

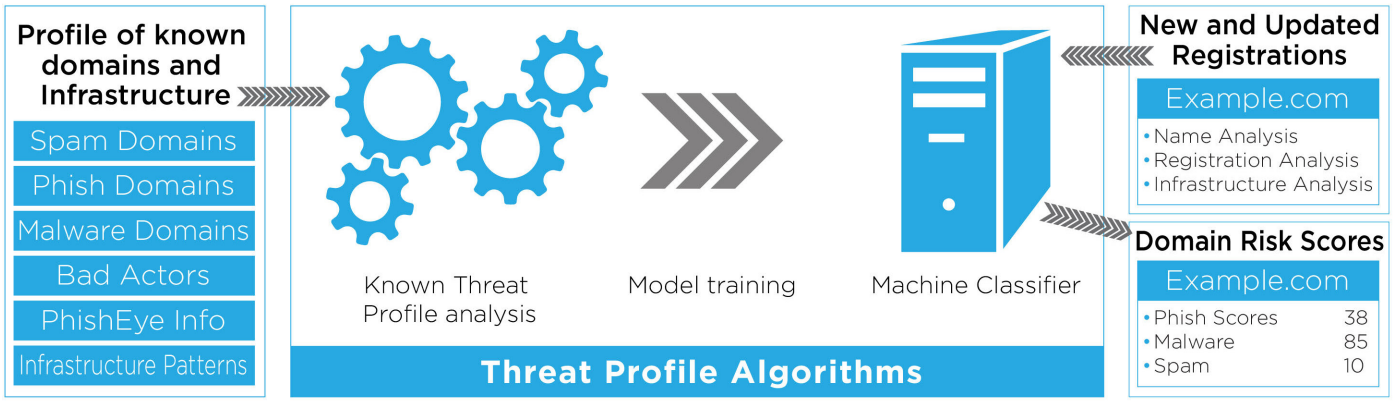


*Figure 1*

## DomainTools Threat Profile

Figure 1: An overview of the DomainTools Threat Profile, showing how malicious domain data is used to train classifiers which then generate risk scores on new and and updated domain registrations

 | sales@domaintools.com | (206) 838-9020 | www.domaintools.com

DOMAINTOOLS

## What is a Threat Profile?

The DomainTools data science team spent nearly two years analyzing the 315+ million domains in our database and created a set of machine learning algorithms called Threat Profile. It offers security professionals a view into the mindset of threat actors: how they determine which domain names to register, and how they set up their malicious infrastructure. From that view, DomainTools created three specific machine learning classifiers—one each for phishing, malware, and spam.

Each classifier is independently engineered, trained, and optimized to find domains registered with malicious intent. The 'seed' for the analysis is the same high-quality blacklist information as the Proximity component of the Risk Score, as well as our extensive domain profile and DNS databases. From these sources, our data scientists identified important domain features against which to train our classification models. Each model is repeatedly tested and optimized over time to validate its accuracy.

These classifiers analyze all new and updated domain registrations, generating scores indicating the strength of the 'signal' that a given domain has malicious intent. A high score does not guarantee badness; the actor may register many domains but only end up using a few. Regardless, the Threat Profile classifiers are designed to find all such domains registered by threat actors whether or not they ever become weaponized.

***Note: Unlike observation-based threat feeds that detect dangerous behavior on the compromised domains, Threat Profile does not not seek to identify legitimate domains that have been compromised. Threat Profile specifically targets domains we believe are registered with malicious intent and for this reason, it only applies scores to domains of an age of under 30 months.***

## Classifying Domains

- Each of the three state-of-the-art Threat Profile classifiers was developed and tested via the following process:
- Create training and test datasets using curated blacklist data and our Whois and DNS databases
- Use our extensive domain knowledge of bad actors, our expertise in cybersecurity TTPs, and detailed analysis of our data to determine which features* are most useful for identifying malicious intent
- Run grid searches using our machine learning infrastructure over different sets of features and tunings to optimize our classification models
- Compare model accuracy on the test dataset using standard classification metrics

* A ***feature*** is machine learning terminology for an intrinsic property of an item which is used to train a classifier or used to classify an item. Raw information about an item is not useful; it needs to be encoded for a computer to understand it. This encoding is the feature. For example, to train a classifier to predict a person's age, height as measured in inches could be a useful feature. That feature would be encoded as a number to represent height. However, height obviously does not reliably predict age by itself; it is one of a set of multiple features that a human age classifier might use to develop a reliable prediction. When a classifier does training, it looks at the set features for each item (person, or domain) and learns from the patterns it finds.

 

Ohe Threat Profile classifiers use features from three categories of data:

- The domain name itself, including TLD
- Domain registration information
- Domain infrastructure information

Different features can be more or less important when predicting phish, malware, or spam intention. As a concrete example, a hyphen in a domain name, such as "com-online-today[.]test", turns out to be an important signal for identifying phishing domains, while it is not nearly as important for identifying spam domains.

For each classifier, we look for discriminatory features in two ways. First, we leverage DomainTools' internal expertise in cybersecurity and domain registrations. For example, many of the features used in the phish classifier come from research behind the creation of PhishEye. Secondly comes true data science, which looks for correlations and patterns among domain metadata. This analysis runs deeper than just the characters in the domain name or TLD; it looks at how and when a domain was registered, and analyzes the infrastructure used to host the domain.

## Train, Test, Repeat

For machine learning on any large and evolving data set, it is vital to continually experiment with different combinations of features. To that end, the data scientists created a robust infrastructure to rapidly deploy and test changes to the features and classifiers. Using this model, we can run not just a few, but hundreds of classifier experiments on our cluster at once. This enables the team to find interesting interactions between features, and improve the models. Specifically, we randomly sample domains to be in either a training or a test dataset, and then perform k-fold cross-validation over models built with the training dataset. This helps ensure that the models are not brittle or overly sensitive to the training data. K-folds are useful enough that we built them into the model.

DomainTools uses a standard set of accuracy metrics to evaluate our models. Some metrics measure the classifier's overall performance, and others measure its performance at a given threshold. We evaluate against the withheld test datasets. Our metrics include:

- Receiver-Operator Characteristic ("ROC") Curves
- Precision-Recall ("PR") Curves
- Precision, Recall, and the F1 Score, at given thresholds

For both ROC and PR curves, it is common to look both at a visualization of the curve as well as the area under the curve (AUC). The higher the AUC, the better the classifier is performing, with 1.0 being "perfect". The F1 score is the harmonic mean of precision and recall, and thus takes both false positives and false negatives into account. This is more robust than precision or recall alone, making it harder to achieve a high score. It is ideal for the rigor demanded in accurate domain classification. As with the AUC, the F1 score ranges from 0.0 to 1.0.

 

## Peeking Under the Hood

As an indication of Threat Profile's accuracy, here is a look at the metrics for the Phishing classifier as of a point just before the general release of Domain Risk Score:

Table 1 shows some summary metric scores for our Phishing Threat Profile classifier.

| Summary Metrics for Phishing | |
|---|---|
| PR AUC | 0.968 |
| ROC AUC | 0.971 |
| F1 Score | .914 at threshold .44 |

*Table 1: Summary Metrics for Phishing*

The classifier for the Phish Threat Profile returns a raw score between 0 and 1, where 0 means not "phishy" at all and 1 means completely "phishy". To compare the classifier's score against the test dataset, the team selects a threshold, typically 0.5, and does a "cut". Everything below the threshold is considered a 0 (not phishy), and everything above it is a 1 (completely phishy). For this instance of Phish, we set our threshold at 0.46.

Figure 2 shows how the Precision, Recall, and F1 scores for Phish vary as we adjust the Threshold parameter from 0 to 1. In the figure, one can trace the tradeoffs between optimizing a classifier for precision versus recall: as one falls the other increases. Happily, for a broad set of thresholds the Phish classifier generates high F1 scores. This implies that most of the raw classification scores are not near 0.5, but rather towards the two ends of the spectrum, which gives high confidence in the quality of classifications.

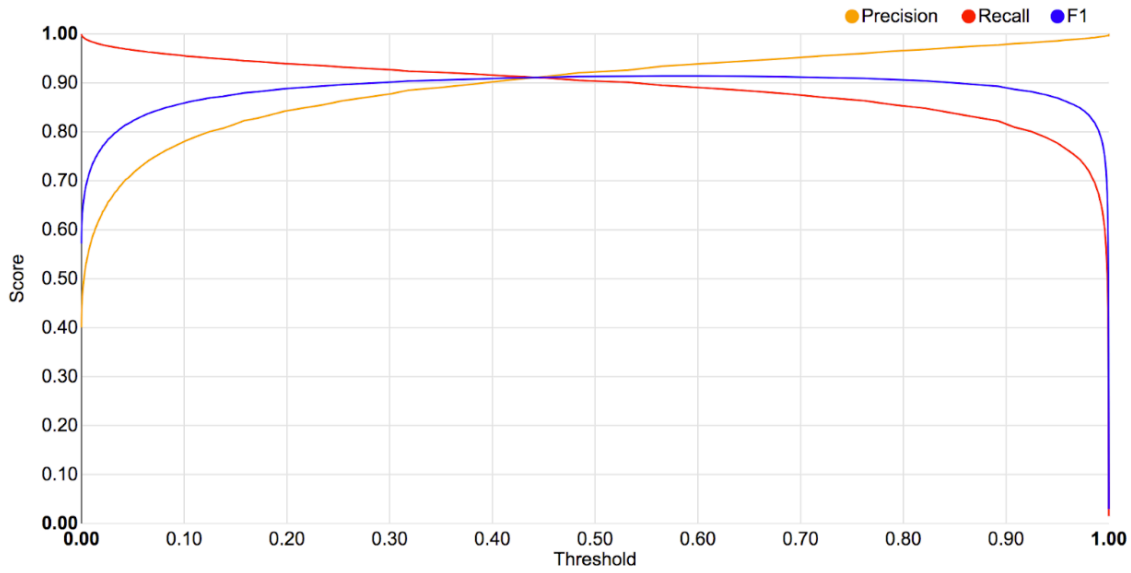


*Figure 2: Precision, Recall, and F1 scores for Threat Profile Phish by Threshold. The x-axis is the threshold, and the y axis is the metric score.*

The Malware and Spam Threat Profile classifiers show even better performance, with F1 scores near or exceeding 0.9 and ROC AUC scores above 0.95.

 

## Interpretation and Use

Threat Profile is a risk score and intended to be used as part of existing threat intelligence processes. As mentioned earlier, one can think of domains with a high Threat Profile score as belonging on a "domain watchlist," meaning they could become weaponized anytime within the next 18 months. Depending on how severely we score the domain and your organization's risk tolerance, it may be desirable to take different actions—everything from flagging their appearance in a SIEM or in server logs, to blocking the domains outright.

The Threat Profile score format is similar to the Proximity format, following a 0 to 100 scale. The higher the Threat Profile score, the more likely the domain was registered with malicious intent:

- 0, domain is whitelisted
- 50+, suspicious
- 70+, our recommended threshold for indicating malicious intent
- 90+, strong confidence in near-term weaponization
- 100, domain is already blacklisted

Threat Profile combines the results from the three independent classifiers together to create one composite score. In the Risk Evidence API endpoint, as well as in Iris, we provide supporting evidence, which summarizes how the classifiers fed into the score for a given domain. Threat Profile is provided in conjunction with with the Proximity score to help an analyst understand the kinds of threats appearing on the protected network–Proximity to identify domains closely related to known malicious activity, and Threat Profile to identify domains with malicious intent before they can be weaponized. The overall Risk Score represents the strongest signal from Threat Profile or Proximity.

A Note About Dormant Domains: Not every domain registered by a bad actor will be weaponized. Many will sit dormant until their registration period ends. The goal of Threat Profile is to find all domains registered with malicious intent, even if they remain dormant. From a classification perspective, these domains are not "false positives" but rather "future positives," because the classifiers indicate that they have the potential to become weaponized at any time.

Every security team faces a tradeoff between providing users access to Internet resources, and protecting the network from threats. DomainTools believes that watching and/or blocking domains flagged by Risk Score is an effective way to isolate potential threats while minimizing the impact to trusted users and customers.

## The DomainTools Advantage

DomainTools' dedicated R&D and Data Science team continually monitor changes in our DNS and domain profile databases, and evaluate new blacklisted domains to detect changes in threat actor behavior and update our models accordingly, taking advantage our ability to rapidly iterate. This infrastructure is just as important as the Risk Scores themselves–it means DomainTools can keep making accurate predictions in the future, no matter how threat actors change their tactics.

To run an evaluation of Domain Risk Score in your environment via the API, as an add-on for Iris, or in a feed of all high-risk domains, contact **sales@domaintools.com**.

 