Proxy in a Haystack: Uncovering and Classifying MFA Bypass Phishing Attacks in Large-Scale Authentication Data

Rebecca Lynch[1], M.C.S., Lauren Saue-Fletcher[2]

[1]*Sr. Data Scientist, Security Research, Duo Security, Cisco Secure*
[2]*Graduate Student, Stanford Empirical Security Research Group, Stanford University*

Abstract

While phishing has long been a prevalent threat against authentication systems, a recent gain in popularity of OSS reverse-proxy kits has made detection and prevention of phishing attacks increasingly difficult. Open-source tools such as evilginx are capable of not only phishing credentials and passcodes, but proxying an entire multi-factor authentication (MFA) flow and all associated cookies. In this scenario, the user sees an expected login prompt from the MFA provider, proxied through the attack server, while the MFA provider sees what appears to be a valid login session simply originating from a different IP address. To the authentication provider, the IP of the attack server is often the only apparent difference between a malicious and a benign authentication. This, coupled with inaccuracies in IP geolocation databases, highly variable user behaviors, ISP IP shuffling, benign VPN usage, and a severe imbalance between benign and malicious authentications, limits traditional server-side ML detection capabilities. Using data from Duo Security, a large authentication provider, we apply point-in-time DNS data to authentication records to identify domains corresponding to the source IP address of the client at the moment of access. We then applied targeted URL and behavioral filtering to identify likely attacker-owned domain-IP pairs. We analyzed authentications from these IP addresses to provide new insights on MFA phishing attack signatures. With this newly uncovered set of labeled malicious authentications, we test a variety of classification approaches in the detection of MFA bypass attacks. We demonstrate the benefits of threat-informed data mining in true positive sample generation, as well as the performance and usability tradeoffs of multiple classification methods in the server-side detection of MFA bypass attacks. These classification techniques applied on newly labeled phishing authentication data are then shown to out-perform unsupervised methods in the identification of malicious authentications.

Keywords

multi-factor authentication, phishing, threat detection,

1. Introduction

1.1. Terminology

We use the term **"access device"** as the device initiating an authentication, **"authentication device"** as the (optional) device approving the authentication, such as a mobile phone approving a Push request, and **"user"** as the end user attempting to authenticate. Users belong to an

CAMLIS'23: Conference on Applied Machine Learning for Information Security, October 19–20, 2023, Arlington, VA
*Corresponding author.
beccalyn@cisco.com (R. Lynch); laurensauefletcher@stanford.edu (L. Saue-Fletcher)
beccalynch.com (R. Lynch)

"customer" or organization which has configured an authentication provider to protect one or more **"applications"** that a user accesses.

1.2. Phishing Attacks

Phishing attacks remain the most prevalent methodology leveraged by bad actors; a recent estimate[1] by CISA approximates that 90% of all recent cyber incidents began with some sort of phishing vector. A 2023 analysis by ZScalar[2] showed that phishing attacks witnessed a staggering increase of 47.2% in 2022. In that time, an estimated $52 million USD was lost as a direct result of phishing attacks[3], though that number is likely higher as only approximately 2.1% of phishing attacks are actually reported[4]. Traditional phishing requires a non-trivial amount of effort to execute, often requiring the creation of a fraudulent website. The attacker then lures users into sharing credentials and, if needed, MFA tokens. In the case of MFA-targeted phishing, these attacks often focus on "phishable" authentication methods, such as One-Time Passcodes (OTP) sent via SMS or mobile apps. While providing more security than no MFA, these MFA methods can be trivially phished. MFA codes typically remain valid until used, allowing the attacker to obtain them from the user via their phishing site and replay them to the authentication provider for unauthorized access.

1.3. Reverse-Proxy MFA Bypass Attacks

While prevention of phishing attacks has improved as more secure MFA factors are increasingly adopted, a method of phishing has been developed that bypasses most MFA factors altogether[5]. With this adversary-in-the-middle (AitM) approach, traffic between the authentication provider and the victim is directly proxied through an attack server, significantly reducing the effort required by an attacker, as spinning up a custom phishing site is no longer necessary. Tools like evilginx are entirely open-source and offer pre-configured proxy kits for a number of popular sites such as Facebook, Twitter, Outlook, and Paypal. These tools proxy the entire login flow that mirrors the user's expected experience with almost no setup required by the attacker. Upon a victim accessing the attack server via a phishing URL, the server negotiates an SSL connection with both the victim and the authentication provider, giving the attacker decrypted access to all credentials, MFA codes, and cookies shared between the two. The URL is typically the only perceptible difference for an end user. This difference is generally of limited use, as 38 of 70 surveyed users in our own simulated internal attack reported that they did not check the URL prior to clicking the link. Phishing resistant authentication methods, such as FIDO2, defend against reverse-proxy phishing attacks, however these authentication methods are often more difficult to deploy in practice.[6]

As an authentication provider, we find these types of attacks to be difficult to detect, as the attack server proxies all information including user-agent strings and OS telemetry throughout the login. Attackers can even subvert client-side detection via the ability to inject and overwrite the Javascript loaded by the client. Because of this, our server-side detection capabilities via traditional machine learning methods are extremely limited. Often, the only perceptibly different signal is the source IP address of the login attempt which will belong to the attack server rather than the user themself. Attempts to classify attack instances are generally limited

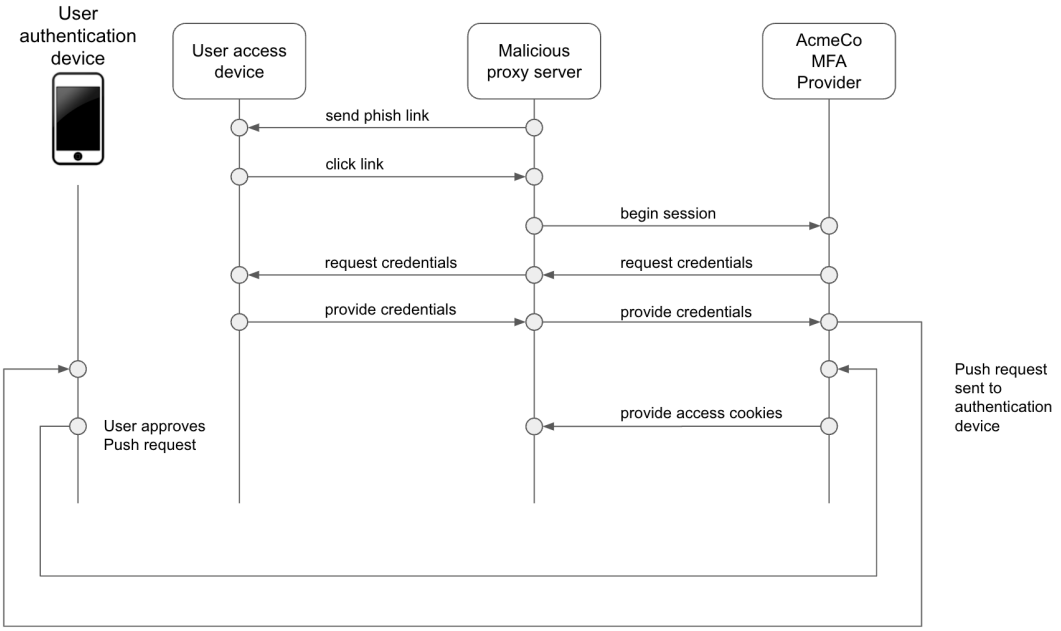

Figure 1: An overview of hew reverse-proxy AitM servers bypass Push-based MFA

by a scarcity of labeled data; labeling true positive attacks is exceedingly difficult, as there are a number of benign reasons for users to authenticate through proxy services or utilize different networks throughout their standard authentication behavior. Additionally, the sheer volume of authentications and variety of potential implementations of this style of attack has previously limited our detection capabilities to unsupervised anomaly detection methods.

1.4. Detection Improvements via DNS Data Integration

With these limitations in mind, we propose the improvement of server-side detection of AitM phishing attacks with the integration of DNS intelligence relating to observed access device IP addresses. Available signals for detection are limited to the IP address of the device initiating access with the authentication provider, the IP address, if relevant, of the authentication device approving the authentication such as a mobile phone, subsequent IP metadata, as well as telemetry received from request headers throughout the authentication. We employ a system in which point-in-time DNS information is used to identify IP addresses suspected of running these reverse-proxy phishing servers. This approach is based on an understanding of the attack topology – to employ this attack, a proxy server must be configured with a domain and valid SSL certificate in order to effectively phish a user. This same server is (in most cases) the server that initiates the connection with the authentication provider. This understanding allows us to identify access device IP addresses corresponding to registered domains as a means of more targeted threat identification. Once these domains were filtered and vetted, we then labeled authentications deriving from these access IPs as phishing attacks. These labeled instances

can then be used to improve detection capabilities with the introduction of supervised ML approaches, providing a threat-informed path beyond strictly unsupervised methods.

2. Methodology

2.1. Threat-Informed Data Labeling via DNS Data Integration

Spanning Duo authentication log data from 2023-05-10 to 2022-05-24, we identified 22,280,355 unique access device IP addresses. Our first goal in identifying attack servers was to find IP addresses associated with potential phishing domains; if a user is currently experiencing an AitM attack, it's likely that the phishing URL they accessed is mapped to the IP address seen in our data, because the attacker's server will be the one initiating the authentication with Duo Security. To do this, Farsight Security's DNS query data was joined against this authentication data to identify domains that were associated with an authentication access device IP. Roughly 300,000 unique domains were found to map to access device IPs within that window. To narrow this, we first considered common attributes of phishing URLs such as the existence of phish "hint" words (login, cash, quick, auth, etc.), common brand names or brand misspellings in the URL[7], as well as the presence of repeated characters, symbols, or misleading TLDs such as google[.]com[.]uz. Further work was done to determine the age of the domains and associated autonomous system number (ASN) in filtering out legitimate domains. Domains older than one year or owned by trusted organizations such as educational or government agencies were ruled out. These filtering measures led to the identification of roughly 300 potentially suspicious domain-IP pairs. We then analyzed the authentications originating from these IPs to determine the likelihood of malicious activity from these domain-IP pairs. IPs with regular usage by a consistent set of users were ruled out as either belonging to legitimate proxy services that are typically used in Zero Trust networks, or personal proxies used to subvert organizational or government censorship. While not inherently benign, this activity is not likely associated with phishing and therefore not relevant in our identification of possible AitM servers. Further filtering was done to remove IPs for which users had extensive history in our data, likely indicating a home network on which they host a personal domain. With this filtering, we identified 14 domains matching both suspicious URL markers and authentication behavior that would suggest a phishing campaign – these domains are listed in the Appendix. These behaviors include one-off authentications across multiple users and ASNs associated with known hosting providers that offer free or cheap web hosting. A number of these sites were still accessible at the time this was written, their alleged purposes ranging from mail/package reception, moving services, Starbucks ordering, and, most prevalently, tutoring and educational resources for college students. Of the 14 domain-IP pairs identified, there were over 25 impacted users, accounting for 77 authentication attempts within the analyzed two week window. 61% of the discovered phishing authentications impacted users at educational institutions, with financial services and manufacturing representing the next most common vectors at 14% and 6% respectively.

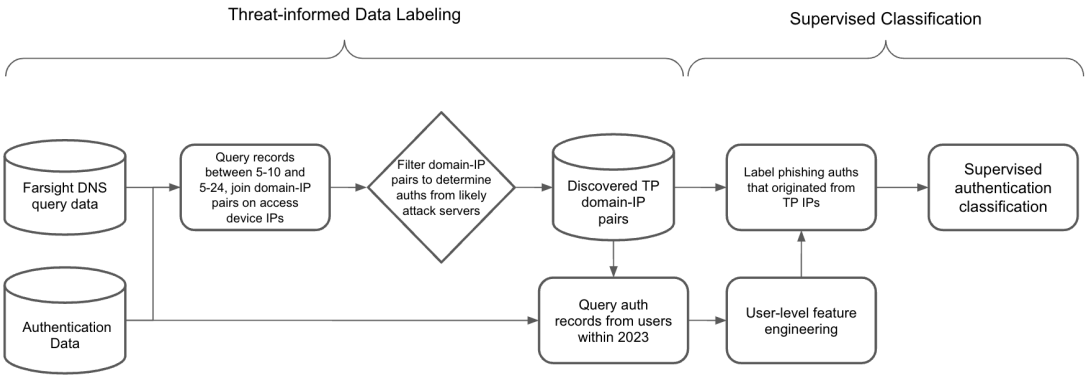

Figure 2: An overview of targeted data labeling via DNS data integration and subsequent feature generation and classification

2.2. Feature Engineering

For the users impacted by these suspected AitM servers, we pulled the entirety of their authentication history in 2023. Authentications within the 5/10 - 5/24 window originating from the phishing IPs were labeled as **phish** authentications, while the rest were labeled as **benign**. The resulting dataset contained 77 phishing authentications and 12,561 benign authentications. Real-world authentication data is highly imbalanced; for every billion authentications that we process, confirmed true positive attack reports are in the single digits. For this reason, we operated within the constraint of this imbalance and chose to not generate synthetic malicious samples or down-sample the benign class. We generated rolling features based on a targeted understanding of the likely taxonomy of a phishing attack. Many of the features are a computed likelihood based on previous authentication data for each user, implemented as the percentage match of a feature or feature pair over a user's 90 day successful authentication history. Features were largely generated at the user level to allow for a generalized classification approach. As each user behaves differently, even within the same organization, it is necessary to consider features as they pertain to an individual's history, rather than make generalizations about suspected attack behavior and introduce unnecessary bias. These rolling probabilistic features include (1) the access device browser type, (2) the access device's country and state of origin as inferred from MaxMind's geo-IP dataset, (3) the pairing of the access device IP's ASN and the application being accessed, (4) the pairing of the MFA factor used and the application being accessed, and (5) the pairing of the access device operating system and the application being accessed. In short, rather than classify on these features themselves, we classify on the user's probability that their authentication would have each feature value. We additionally included boolean features to indicate (6) whether an access device browser's version had decreased since the prior successful authentication, and (7) whether the access device carrier (e.g. Comcast, Amazon, DigitalOcean) associated with the IP has changed since the prior successful authentication. We additionally incorporated known effective features from prior risk-based assessment

work, including (8) whether the access device ASN and (9) IP address are novel within a user's organization, (10) whether the access device IP has been seen by a different organization within a 24 hour lookback, and (11) the distance between the access device location of the current and prior authentication as well as (12) the average distance between the access device location of the current and last ten authentications.

2.3. Classification of Phishing Authentications

We used both XGBoost and LightGBM to classify these malicious authentications, chosen due to the cardinality and extreme imbalance within our sample dataset. These models were tuned to optimize the F_1 score. The F_1 score is defined here, using p = precision and r = recall.

$$F_1 = \frac{2 * p * r}{p + r} \tag{1}$$

We selected F_1 as the performance metric due to the data imbalance – when working with imbalanced classes, we must optimize for precision (proportion of flagged records that are correctly identified as malicious) and recall (proportion of truly malicious records that are correctly flagged). The tuned parameters for each model can be found in Appendix B. Additionally, an unsupervised Isolation Forest model was employed as a benchmark, as similar detection methods are currently used at Duo due to the previously described limitations in labeled data. The contamination rate for the IF model was set at 0.01 to properly represent the rate of imbalance in the training dataset.

3. Results

3.1. Feature Correlation with True Positive Phishing Attacks

Generated features were assessed against benign and phishing authentications to better understand the signals that separated the identified true positives. The full set of visualizations for these features are shown in Appendix A. Of the probabilistic features, the probability of seeing a given ASN and application pair had one of the highest levels of separation between classes, with location probabilities also showing discernible differences.

Among boolean features, we see the most notable difference between classes for the features involving the distance between authentications, with the majority of phishing authentications having a > 100 mile distance from the last successful authentication, compared to only 20% of legitimate authentications.

Table 1
Classification Results

Model	Recall	Precision	Accuracy
XGBoost	0.63	0.02	0.64
LightGBM	0.61	0.05	0.81
Isolation Forest	0.06	0.08	0.98

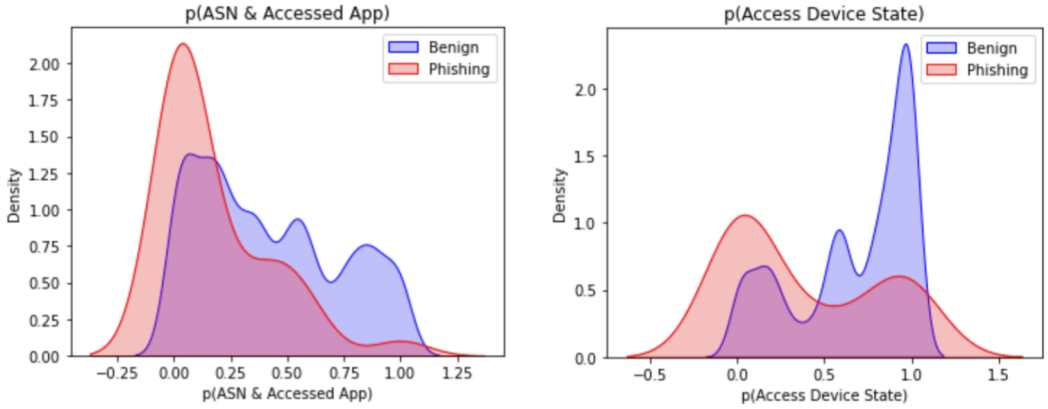

Figure 3: Density curves showing distributions of probabilities for phishing vs. benign authentications, for ASN + application and access device state probabilities respectively

3.2. Supervised vs Unsupervised Classification of Phishing Authentications

Precision, recall, and accuracy are shown below for each approach. It is important to note preemptively that these metrics were measured purely against the generated user-level time series data and **do not take into account features that would objectively improve both accuracy and precision.** These include policy measures that organizations using Duo commonly employ, including allowlisted networks, remembered devices, and secure FIDO2 factors. In a real-world application, authentications meeting these criteria would not be flagged as malicious regardless of the time-series features used here.

4. Discussion

4.1. Label Generation

It is essential to maintain a high-confidence labeled dataset in any classification problem. This is especially crucial in the realm of cybersecurity, as the risk posed by false negatives (failing to identify a malicious authentication) can lead not only to financial loss but significant impacts to victims' lives. The risk associated with false positives (incorrectly marking benign authentications as malicious) is also moderate, as our services are used to protect critical applications including medical software, university portals, and software reliability instructure. In the authentication space, the large volume of highly cardinal data is further compounded by the amount of variability within user-level data, making identifying malicious authentications incredibly difficult. In this case, even with the augmentation of our authentication data with DNS query information, significant domain knowledge was necessary to uncover high confidence true positives. When looking for instances of this particular attack, however, our approach to data augmentation significantly reduced our search space from 800 million authentications originating from 20 million unique IPs, to 300,000 IPs with corresponding domains, to only 300 IPs with highly suspicious domains. We plan to improve upon this filtering method and

design a real-time system by which this data can be integrated into our detection systems and allow us to continue to build a set of high confidence true positives. Features found in the previous section to be highly correlated with phishing attacks can be used to aid in further threat research as a means of narrowing this search for threat researchers seeking to identify malicious behavior when DNS information may not be available.

4.2. Limitations of Classification Techniques on Authentication Data

Classification on authentication data is inherently difficult. The vast majority of authentications in our dataset are not malicious. While our dataset was limited by a small count of confirmed true positives, the nature of authentication data would likely lead to similar performance numbers even with the presence of more identified phishing IPs for several reasons. First, the efficacy of the benign authentication labels is generally unknown. While we can reasonably attest that our labeled true positives are from malicious attack servers, our confidence that the labeled benign authentications are truly benign is generally lower. This is likely to lead to a degradation of precision, as we cannot necessarily ensure that all misclassified "attack" authentications are truly misclassified. Second, users in general exhibit many legitimate behaviors that may appear as malicious or anomalous activity. Behavior that is normal for one user may be indicative of a malicious authentication for a different user. The behavior among individuals using Duo's authentication services varies greatly: the average user in our analyzed data utilized 4.5 unique network carriers, 18 unique IP addresses, and 2.2 distinct operating systems over a two month period. User behavior also varies seasonally. Many users are university students that exhibit a dramatic shift in activity at the start and end of the school year, both in terms of the features of the behavior (different access device types, locations, VPN utilization), and in authentication volume. With these limitations in mind, the intention of this work is not to propose novel or perfect ML methods of detection, but rather to describe the application of threat-informed data filtering in providing a path forward when dealing with the detection of an otherwise imperceptible attack. That said, the supervised classification methods we were able to use as a direct result of targeted threat-informed data labeling shows a profound improvement in detection recall over currently employed unsupervised methods. We intend to use these findings to develop a fully integrated DNS-aware authentication classification system that can extend these methods and, with the aid of human label verification, continue to build a set of informed true positive malicious authentications and improve our automated detection capabilities.

References

[1] CISA, Stop ransomware, 2023. URL: https://www.cisa.gov/stopransomware/general-information.

[2] D. Desai, R. Hedge, E. Laufer, J. Wang, 2023 phishing report reveals 47.2% surge in phishing attacks, 2023. https://www.zscaler.com/blogs/security-research/2023-phishing-report-reveals-47-2-surge-phishing-attacks-last-year.

[3] Federal bureau of investigation internet crime report, 2022. URL: https://www.ic3.gov/Media/PDF/AnnualReport/2022State/StateReport.aspx.

[4] C. Baron, 28% of bec attacks opened by employees, new data shows, 2023. URL: https://abnormalsecurity.com/blog/28-of-bec-attacks-opened-by-employees.
[5] B. Toulas, Mfa adoption pushes phishing actors to reverse-proxy solutions, 2022.
[6] M. Kapko, What is phishing-resistant multifactor authentication? it's complicated., 2022. URL: https://www.cybersecuritydive.com/news/phishing-resistant-mfa/633703/.
[7] A. Moraru, P. R. Donahue, Top 50 most impersonated brands in phishing attacks and new tools you can use to protect your employees from them, 2023. https://blog.cloudflare.com/50-most-impersonated-brands-protect-phishing/.

A. Feature Separation

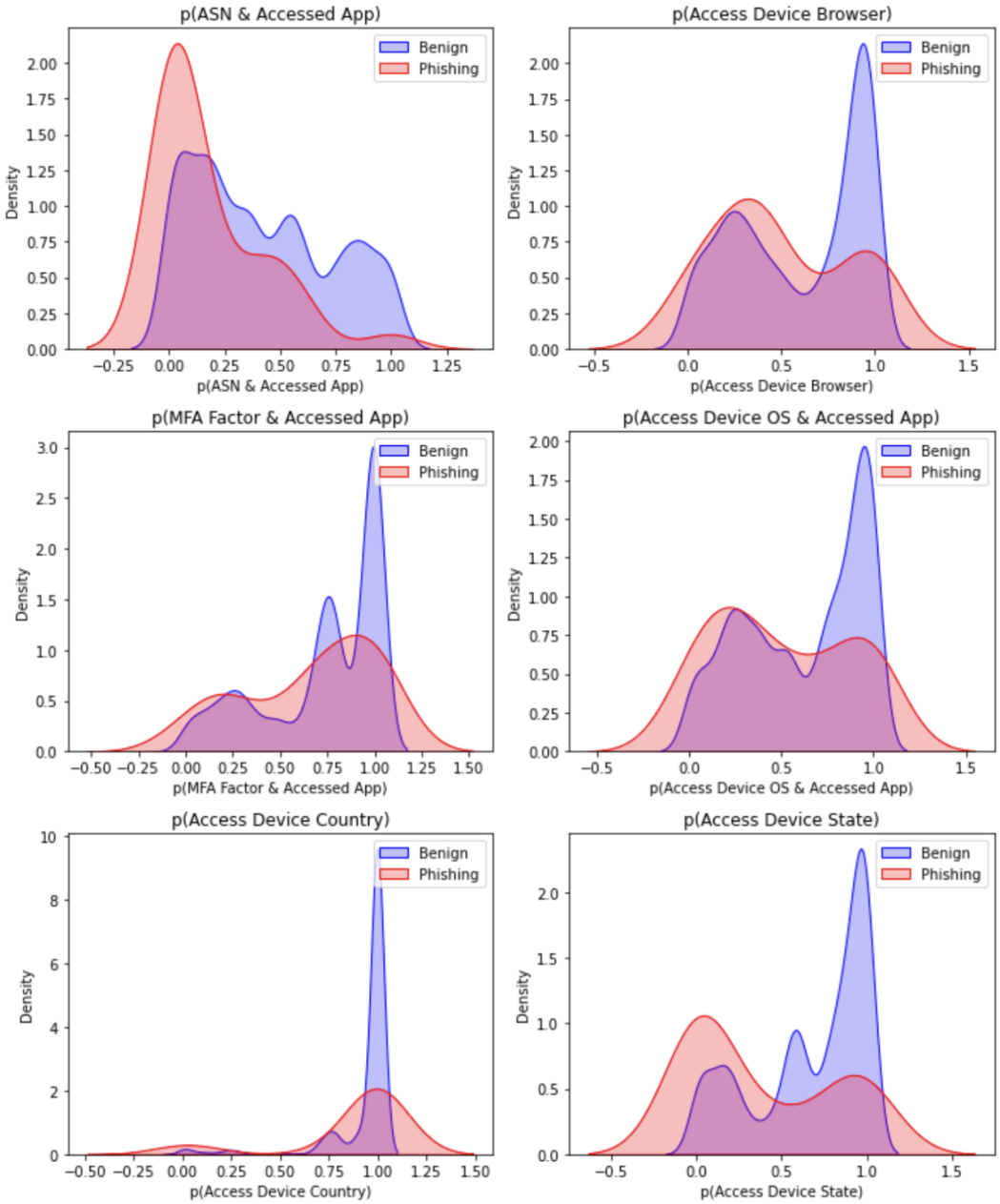

Figure 4: Density curves showing distributions of probabilities for phishing vs. benign authentications.

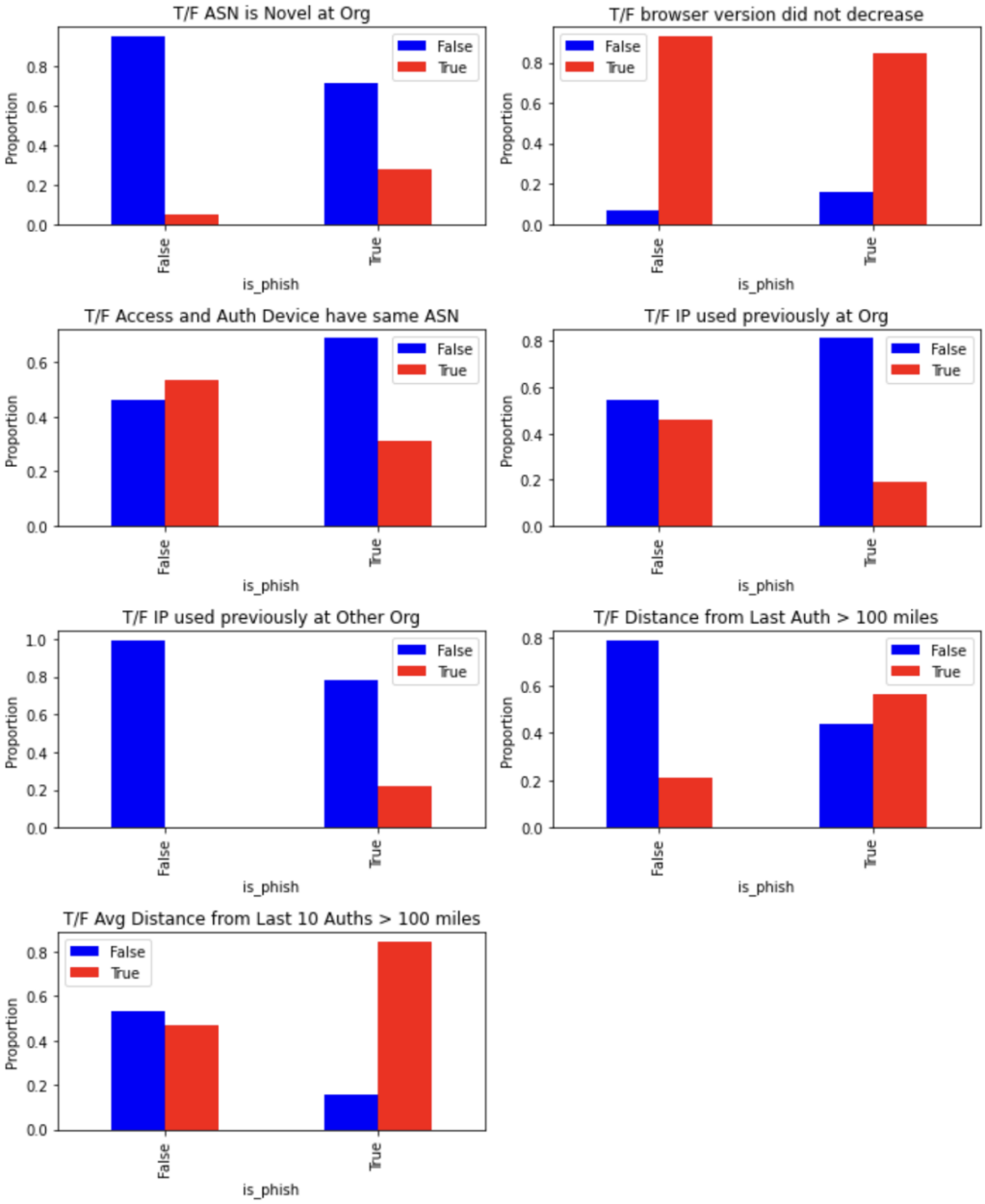

Figure 5: Proportions of boolean values for phishing vs. benign authentications.

B. Domains

- package-usps[.]us
- *[.]zetlandcapitals[.]com
- criteriacorp[.]microsoftonline[.]app-account-127[.]cloud
- volvo[.]microsoftonline[.]app-account-140[.]cloud
- cbeyondata[.]microsoftonline[.]app-account-126[.]cloud
- b9746927-a325-5d2d-7f91-ca0105ac5f52[.]cnnic[.]rip
- t3[.]freegradely[.]xyz
- starburkx[.]com
- gooduugfdhgf[.]click
- clientedesco004[.]descobrresgate[.]com
- dvfffpyvl[.]mom
- lswj35[.]suporteswr[.]com
- uiuvjfkkge[.]buzz
- wwwofc[.]getgoingmove[.]com