Computer Fraud and Security
ISSN (online): 1873-7056

Validation of a Phishing Risk Model in the Banking Sector based on Historical Data

Jerzy Dorobisz

Affiliation e.g. Institute of Information Technology and Cyber-security, Faculty of Cybernetics, WAT,
2 Gen. Sylwestra Kaliskiego St., 00-908 Warsaw

email jerzy.dorobisz@wat.edu.pl

Due to the growing digitisation of banking services, phishing remains one of the most significant threats in the banking sector. This paper presents a time-aware, standards-aligned framework for validating a phishing risk model using historical data while addressing class imbalance, calibration, and cost-sensitive thresholds. We outline data sources, privacy safeguards, feature engineering, and modelling families (regularised logistic regression, gradient boosting, random forests), together with blocked time-series validation and a cold hold-out. We emphasise reporting with PR-AUC, ROC-AUC, precision/recall@k, and calibration (Brier score), and connect model quality to operational value in SOC/ERM processes. An **illustrative** results section shows how precision–recall and calibration guide threshold selection and reduce analyst workload at fixed recall. The approach supports continuous improvement of anti-phishing defences and provides an auditable protocol for model governance.[1]

Keywords: phishing; electronic banking; risk assessment; model validation; historical data; machine learning

1. Preliminary analysis of the phishing risk problem in the banking sector

In the context of validating a phishing risk model, it is crucial to understand the network structure of a bank, which is the first line of defence. Modern banks use a multi-layered architecture, including elements such as:

- Network segmentation: This involves creating separate zones (e.g. DMZ for public servers, internal network for transaction systems) to minimise the risk of an attacker moving around.

- Access control mechanisms: These include firewalls, IDS/IPS systems, and advanced anti-spam/anti-phishing filters that analyse message content for phishing indicators such as invalid URLs, suspicious attachments, or domain spoofing attempts.

- Anomaly detection systems: These monitor network traffic for unusual patterns (e.g., mass data transfers to external IP addresses) that may indicate information leakage following a successful phishing attack.

- Cloud integration: Hybrid environments require additional security measures, such as CASB (Cloud Access Security Broker), to prevent cloud services from being used to spread malicious links.

A key challenge is to protect user touchpoints, such as employee email inboxes, which are the main vector for phishing attacks. A large proportion of attacks use crafted emails with PDF/Office attachments containing malicious code.[2]

2. Identifying critical assets

To effectively defend against threats, the first step is to identify the most sensitive resources used by banks. These include:

- Databases storing card numbers, authentication data (e.g. 2FA tokens) and customer transaction history. Phishing attacks often target employees who haveaccess to these systems, using social engineering techniques (e.g. fake transfer orders).

- Certification centres and electronic signature mechanisms. Phishing targeted at the administrators of these systems can lead to the compromise of the entire infrastructure.

- Personal data such as names, addresses, and PESEL numbers – used for identity theft or targeted phishing. Phishing attacks use personal data to personalise messages, increasing their credibility.

- Customer communication channels, such as official websites and mobile applications. *Typosquatting* (registering domains similar to those of banks) or *spoofing* (impersonating trusted sources) attacks can lead to the theft of customer login details.

- Communication systems such as Microsoft Teams. Phishing attacks using fake notifications or links to "urgent documents" are becoming increasingly common.[3]

3. Main reasons for failure to defend against attacks

The weakest link in phishing attacks remains the human factor. The main reasons why attacks are successful are:

- Configuration errors in spam filters, allowing phishing messages to reach inboxes.

- Outdated systems without security updates, vulnerable to exploits delivered by attachments.

- Insufficient training – not all banks conduct regular phishing attack simulations for their employees.[4]

4. The most important safeguards against phishing attacks

The consequences of potential attacks can be disastrous for a bank. They cause both economic and reputational damage. A number of tools and technologies have been developed to defend against this. The most commonly used are:

- **SSL/TLS encryption**

The basis for secure communication in electronic banking is the use of SSL/TLS protocols. TLS (*Transport Layer Security)* provides three key functions: data encryption (confidentiality protection), server identity verification (protection against phishing and redirection to fake websites) and integrity of transmitted information (protection against data modification during transmission). Modern banks should use the latest versions of TLS (at least 1 .2, and preferably 1 .3), regularly renew and verify SSL/TLS certificates, and enforce HTTPS connections on all digital channels.[4]

- **VPN (*Virtual Private Network*)**

A VPN provides an additional layer of protection by encrypting all network traffic between the user's device and the bank's servers. This is particularly important when working remotely, using public Wi-Fi networks and accessing the bank's administrative systems. A VPN protects against eavesdropping and data interception by third parties, and also allows for secure network traffic segmentation.

- **Multi-factor authentication**

The implementation of multi-factor authentication (e.g. password + SMS, mobile application, biometrics) is now standard in the banking sector and effectively reduces the risk of unauthorised access even in the event of password theft. Multi-factor authentication should be required for all high-risk operations and when accessing administrative systems.[5]

- **Data encryption at rest and in transit**

All sensitive data (transactions, personal data, backups) must be encrypted both during transmission (TLS) and at rest (e.g. encryption of databases, disks, backups). Encryption should be implemented using strong, regularly updated algorithms.[5]

- **Threat monitoring and detection**

The bank should implement advanced monitoring systems (SIEM, IDS/IPS) that analyse network traffic, system logs and security events in real time. This allows for rapid detection of attack attempts and anomalies, and response to incidents.[6]

- **Antivirus software and update management**

All end devices (computers, servers, mobile devices) must be protected by up-to-date antivirus software and regularly updated operating systems and applications.[7]

- **Secure APIs and integrations**

All integrations with external partners should be secured through the use of secure APIs (authorisation, encryption, permission restrictions) and regular penetration testing.[7]

5. Assessing security quality using penetration testing and attack simulations

In the banking sector, where data and system protection is crucial, penetration testing and attack simulations are fundamental tools for assessing the effectiveness of security measures. Their purpose is not only to detect technical vulnerabilities, but also to test the organisation's readiness for real threats, including phishing and social engineering attacks.

Penetration tests are controlled, authorised activities designed to simulate a hacker attack on a bank's infrastructure. Their purpose is to identify weaknesses in systems, applications, networks and security procedures before they can be exploited by criminals. In a bank, these tests help to detect, among other things, vulnerabilities that allow user accounts to be taken over, privileges to be escalated, or financial data to be leaked. We can distinguish the following types of tests:

- Black box – testers have no knowledge of the system, which imitates an external attack.
- White box – testers have full access to documentation and code, allowing for in-depth analysis.
- Grey box – intermediate level of knowledge, e.g. access to a user account.

Phishing simulations are controlled campaigns in which fake emails imitating real phishing attacks are sent to bank employees, but without any actual threat. The aim is to assess staff awareness and response to attempts to obtain information, which is crucial because people are often the weakest link in the security chain.

Penetration tests and simulations usually consist of the following stages:[9]

1) Preparation and determination of the scope of testing

The first and key stage is to precisely define the objectives of the test, its scope and methodology. At this stage, pentesters meet with the client (in this case, the bank) to determine:

- which systems, applications, networks or services will be tested,
- what type of tests will be performed (black-box, white-box, grey-box),
- the objectives of the test (e.g. detection of vulnerabilities enabling account takeover, privilege escalation, data leakage),
- rules and restrictions (e.g. whether the tests may affect the production environment, whether some employees will be informed about the tests).

At this stage, formal agreements (confidentiality agreement) and consent to conduct the tests are also signed, which is necessary for the legality of the activities.[9]

2) Reconnaissance

The reconnaissance phase involves gathering as much information as possible about the target being tested. Pentesters use various techniques and tools to collect data that will help them plan their attacks:

- OSINT (*Open Source Intelligence*) – analysis of publicly available sources, such as bank websites, domain registries, employee profiles on social media, and information about IT infrastructure.
- *Passive reconnaissance* – collecting information without direct contact with the system, e.g. DNS analysis, WHOIS, external port scanning.
- *Active reconnaissance* – direct scanning of networks and systems to detect open ports, services, and software versions that may have known vulnerabilities.

The goal is to build a detailed picture of the infrastructure, potential entry points, and security weaknesses.[9]

3) Scanning and vulnerability identification

At this stage, pentesters use automated tools (e.g. Nmap, Nessus, OpenVAS) and manual techniques to map the infrastructure and detect vulnerabilities such as open ports and services, outdated or vulnerable software versions, weak passwords or configuration errors, and web application errors (e.g. SQL Injection, Cross-Site Scripting). The scan results are analysed in terms of risk and potential ease of exploitation by an attacker.[10]

4) Exploitation

In the exploitation phase, pentesters attempt to exploit the vulnerabilities they have found to gain unauthorised access to systems. Techniques may include brute-force attacks on passwords, SQL injection, exploitation of authentication errors, XSS or CSRF attacks, and exploitation of vulnerabilities in network protocols.

The goal is not only to gain access, but also to simulate real-world attack techniques that could lead to privilege escalation or system takeover.[11]

5) Maintaining access and privilege escalation

After gaining initial access, pentesters try to maintain it for as long as possible, obtaining higher levels of privileges. This stage simulates a scenario in which the attacker wants to maximise damage and access to data. It also tests whether detection and monitoring systems are able to detect such activities.[11]

6) Covering tracks

Pentesters check whether it is possible to remove or hide traces of their activity, which corresponds to the techniques used by real attackers. This stage allows the effectiveness of intrusion detection systems (IDS/IPS) and security policies regarding auditing to be assessed.

7) Reporting and recommendations

After completing the tests, pentesters prepare a detailed report that includes:

- a description of the vulnerabilities detected and their risk level,
- the course of the tests performed and the techniques used,
- the potential consequences of exploiting the vulnerabilities (e.g. data theft, takeover),
- recommendations for repairing and strengthening security measures,
- priorities for corrective actions.

The report is a key tool for the bank's security team, enabling effective planning of corrective actions.[12]

8) Retesting and verification

After implementing the fixes, the bank should commission retests to verify the effectiveness of the corrective measures and ensure that the vulnerabilities have been removed. Retests may include both penetration tests and automated scans.[13]

6. Employee training on security principles

The process of training users in security principles at the bank should be comprehensive, systematic and tailored to the specific nature of the financial sector in order to effectively raise employee awareness and skills in the area of information protection and countering cyber threats. The training process should consist of the following stages:

1. The first stage of the training aims to familiarise employees with the basic principles of data protection and IT security. As part of this module, participants learn about the importance of information security in the bank and their responsibilities in this area. Issues of privacy and personal data protection in accordance with GDPR regulations are discussed, as well as the principles of secure password management and the use of multi-factor authentication. Employees also learn how to use computers, smartphones and networks safely, both at work and at home. This stage is usually carried out in the form of e-learning, enriched with instructional videos, quizzes and knowledge tests that help to reinforce the information learned.

2. The second stage focuses on social engineering threats, which are one of the main causes of security incidents in the banking sector. Participants will learn about the mechanisms of phishing attacks, including their variants such as *spear phishing* and *whaling*. This module covers the analysis of typical characteristics of suspicious emails and attachments, as well as practical tips on how to safely handle suspicious messages. In addition, the role of spam filters and email protection systems is discussed. This stage is often implemented in the form of a webinar with an interactive question and answer session and phishing simulations that allow for practical testing of employee vigilance.

3. In the next stage, participants learn the principles of secure information processing and how to respond to security threats and incidents. The classification and protection of data in the banking environment, the principles of secure use of data carriers and mobile devices, and procedures for reporting incidents and suspected security breaches are discussed. In addition, the importance of backups and backup policies is emphasised. This module can be delivered both in the form of online training and classroom-based classes with practical exercises and case studies, which allow for a better understanding of real-life situations and how to resolve them.

4. The last module covers legal and ethical aspects of working in a bank. Participants learn about banking secrecy rules and their responsibilities in this area. Professional ethics in relations with clients and colleagues, as well as the basics of anti-money laundering and counter-terrorist financing. The training emphasises employees' responsibilities in identifying and monitoring customers and reporting suspicious transactions. This stage is usually delivered as an online training course with interactive materials and knowledge tests to check what has been learned.[14]

Throughout the training process, it is important to use interactive teaching methods such as quizzes, tests, phishing simulations and question and answer sessions, which increase participant engagement and help to reinforce knowledge. Training materials should be available online, allowing for repetition and self-study. E-learning systems also allow for progress monitoring and identification of areas requiring additional support. Upon completion of the training, participants should receive a certificate confirming their acquired competences, which further motivates their engagement.[15]

7. Positive effects of implementing security policies

Reducing the number of phishing attacks brings many significant benefits to the bank, affecting both operational security and customer relations, as well as the institution's financial situation. First and foremost, reducing phishing means a lower risk unauthorised transactions and loss of funds by customers. According to Polish case law, the bank is liable for unauthorised transactions, unless the customer has been grossly negligent, so reducing the number of successful phishing attacks also reduces the potential costs of refunds and complaints, which has a positive impact on the bank's financial stability.

Another positive effect is increased customer confidence in the bank. In an era of growing cyber threats, customers value institutions that effectively protect their data and financial resources. Reducing phishing translates into a better reputation for the bank, which can increase customer loyalty and attract new users to its services. Banks that invest in advanced security technologies, such as multi-factor authentication, spam filters and transaction monitoring, build an image of a modern and responsible institution.[16]

Reducing the number of phishing attacks also improves the bank's operational efficiency. Fewer incidents mean less workload for security and customer service teams, who do not have to spend as much time responding to, analysing and resolving data theft issues. This allows for better use of resources and a focus on developing new services and improving existing security mechanisms.

Furthermore, reducing phishing contributes to better compliance with regulatory requirements and security standards that require banks to manage cyber risk. Lower incident risk makes it easier to meet requirements such as GDPR, NIS2 directives and DORA regulations, minimising the risk of sanctions and financial penalties.[17]

Finally, reducing the number of phishing attacks supports the development of a security culture within the organisation.

Thanks to anti-phishing training and educational campaigns, employees and customers become more aware of the threats and are able to recognise attempts to obtain information more effectively. This, in turn, leads to a lasting increase in the security level of the entire institution.[18

8. Research part: validation of the model on historical data

Research objective and questions

The aim of the research is to empirically validate the phishing risk assessment model in a bank based on historical data and to evaluate its operational usefulness. We answer the following questions:

1. Does the model effectively distinguish between phishing and non-phishing events?
2. How stable are its results over time and across different segments of the organisation?
3. What is **the operational gain** (reduction of damage/false alarms) for the specified decision thresholds?[19]

Data collection and ethics

- **Time frame:** 01.2023–06.2025; **time windows:** *observation window* 14 days, *prediction window* 7 days.
- **Sources:** mail gateway logs, SIEM/EDR, SOC/Helpdesk reports (confirmed incidents), URL/attachment metadata, user telemetry (click-through rate).
- **Ethics/GDPR:** pseudonymised data; no message content – only metadata/features; minimisation of scope, 180-day retention; DPIA for risk profile processing.
- **Labels:** 1 = confirmed phishing, 0 = no incident (excluding unverified reports).
- **Class balance:** note strong imbalance (e.g., 0.5–2.0% positives) and consequences for metrics.[20]

Variables/features (feature engineering)

- **Email:** sender domain reputation, Levenshtein distance of sender from bank domain, SPF/DKIM/DMARC, number and type of attachments, subject entropy, URL density, rare TLDs, *URL shortening*.
- **User/environment:** click history in training campaigns, seniority, organisational unit, time of day/day of the week.
- **Short-term behaviour:** number of "urgent" emails in 24 hours, unusual patterns (from rolling features).
- **Risk aggregates:** reputation indicators (watch lists), number of similar messages in the cohort.[21]

Note: remove leaky features (e.g., fields filled in *after* an incident), standardise/categorise and document the feature dictionary.

Modelling methods

- **Base models:** logistic regression (with regularisation), *gradient boosting* (XGBoost/LightGBM), random forest as a benchmark.
- **Dealing with imbalance:** *class weights* or *focal loss*; SMOTE only for prototypes.
- **Feature selection:** *borutaSHAP* or backward elimination with *variance inflation* control.
- **Explainability:** SHAP (globally and locally) for top features.[22]

Validation scheme

- **Time validation:** *time-series split* (e.g. 5 blocked folds) – without mixing periods.
- **Cold test set:** last quarter/month (not visible in training).
- **Main metrics:** PR-AUC (preferred for rare classes), ROC-AUC, *recall@k* (e.g. @1% and @5% of the highest scores), *precision@k*.

- **Calibration:** *calibration plot* curves, Brier score; possibly *isotonic regression* / *Platt scaling*.
- **Threshold selection:** cost criterion (cost matrix: FN » FP) + Youden J as a reference point.[22]

9. Results (Illustrative Example)

Note: The table uses *synthetic* values to demonstrate reporting format. Replace with your measured results.

Model	PR-AUC	ROC-AUC	Precision@1%	Recall@1%	Brier
LogReg	0.56	0.88	0.42	0.18	0.072
XGBoost	0.71	0.94	0.60	0.31	0.058
RandomForest	0.63	0.91	0.50	0.24	0.061

Template for interpretation: choose the primary model by PR-AUC; report precision/recall at small review depths to quantify analyst workload vs. capture rate; include calibration plots and confidence intervals (e.g., blocked bootstrap). Map improvements to operational KPIs (false-positive reduction at fixed recall, time-to-detection, incident-cost proxies).[22]

7. Discussion

Performance must be reviewed under drift (seasonality, changing lures, macro events) and operational constraints. Higher PR-AUC and precision at small review depths reduce analyst workload and dwell time. Calibration is crucial when scores drive triage or automation. Governance should include a model registry, explainability (e.g., SHAP), auditable documentation, and periodic revalidation (NIST, 2020; ISO/IEC, 2022).[22]

8. Conclusions

A standards-aligned, time-aware validation protocol provides defensible evidence of model quality and operational value. Combining probabilistic metrics, calibration analysis, and cost-based thresholds supports robust decision-making and continuous improvement of anti-phishing defences.

is to protect points of contact with users, especially employee e-mail accounts, which are the main target of phishing attacks. Security measures such as network segmentation, data encryption, multi-factor authentication, advanced monitoring systems and penetration testing are crucial here. Regular employee training is also important, as it increases their vigilance and ability to recognise attempts to obtain information fraudulently. Reducing the number of successful phishing attacks translates into increased customer confidence, improved bank reputation and better compliance with regulatory requirements.

Bibliography

1. Anti-Phishing Working Group. (2025). *Phishing Activity Trends Report (Q4 2024).*
2. CERT Polska (NASK). (2024). *Annual report 2023: Threat landscape in the Polish Internet.*
3. European Union Agency for Cybersecurity. (2024). *ENISA threat landscape 2024.*
4. FIDO Alliance. (2022). *Phishing-resistant authentication: FIDO2/WebAuthn white paper.*
5. Google. (2024). *Safe Browsing / Phishing Protection—Transparency resources.*
6. Internet Engineering Task Force. (2011). *DomainKeys Identified Mail (DKIM) signatures* (RFC 6376).
7. Internet Engineering Task Force. (2014). *Sender Policy Framework (SPF)* (RFC 7208).
8. Internet Engineering Task Force. (2015). *Domain-based Message Authentication, Reporting, and Conformance (DMARC)* (RFC 7489).
9. International Organization for Standardization. (2018). *Risk management—Guidelines* (ISO 31000:2018).
10. International Organization for Standardization & International Electrotechnical Commission. (2022). *Information security, cybersecurity and privacy protection—ISMS—Requirements* (ISO/IEC 27001:2022).

11. International Organization for Standardization & International Electrotechnical Commission. (2022). *Information security, cybersecurity and privacy protection—Guidance on information security risk management* (ISO/IEC 27005:2022).
12. Jakobsson, M., & Myers, S. (Eds.). (2006). *Phishing and countermeasures: Understanding the increasing problem of electronic identity theft.* MIT Press.
13. MITRE Corporation. (2024). *ATT&CK® for Enterprise: TTPs and social engineering/phishing techniques* (Online version).
14. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). *Phishing Websites Data Set* [Data set]. UCI Machine Learning Repository.
15. National Institute of Standards and Technology. (2008). *Technical guide to information security testing and assessment* (NIST SP 800-115).
16. National Institute of Standards and Technology. (2012a). *Guide for conducting risk assessments* (NIST SP 800-30 Rev. 1).
17. National Institute of Standards and Technology. (2012b). *Computer security incident handling guide* (NIST SP 800-61 Rev. 2).
18. National Institute of Standards and Technology. (2017). *Digital identity guidelines* (NIST SP 800-63-3; includes 2019 errata).
19. National Institute of Standards and Technology. (2020). *Security and privacy controls for information systems and organizations* (NIST SP 800-53 Rev. 5).
20. Verizon. (2024). *Data Breach Investigations Report (DBIR) 2024.*
21. World Wide Web Consortium. (2021). *Web Authentication: An API for accessing Public Key Credentials (WebAuthn)—Level 2 (Recommendation).*
22. World Wide Web Consortium. (2023). *Web Authentication: An API for accessing Public Key Credentials (WebAuthn)—Level 3 (Working Draft).*