UNIVERSITY of VIRGINIA
McINTIRE SCHOOL of COMMERCE
Center for Business Analytics

UVA-CBA-0119
Revised May 2, 2019
DRAFT

Exploring a PhishCasting Capability at McAfee

McAfee is a security software company headquartered in Santa Clara, California. The firm delivers an array of digital security tools geared towards consumer and enterprise clients, including offerings for personal computers, mobile devices, and servers. As with all companies operating in the digital security space, anti-phishing solutions have always been one important area of focus. Phishing – a type of semantic attack that exploits human as opposed to software vulnerabilities – is one of the most prevalent forms of cybercrime, impacting over 40 million Internet users every year. Phishing is consistently ranked as one of the top security concerns facing enterprises not only because of the number of employees falling prey to phishing attacks within organizations, but also because brand equity and trust are tarnished when companies' customers are targeted by spoof (i.e., fraudulent replica) websites. An oft-cited statistic in the security industry is that the average 10,000 employee company spends approximately $3.7 million annually combating phishing attacks. Executives at McAfee always felt this statistic significantly underestimated the true monetary cost of phishing.

Developing a Better Mousetrap

When a user clicks on a link (URL), whether it's in an email, web search result, or in a social media post, an anti-phishing tool is used to examine the legitimacy of the web page. If the tool deems the web page to be a phish, a warning appears preventing the page from displaying. However, users have the option of disregarding the warning and proceeding to the potential phishing web page anyway. Many anti-phishing tools exist, including web browser plugins developed by Microsoft, Mozilla, Google, and Apple, and client security software packages offered by Symantec, McAfee, and others. In 2006, McAfee revamped their anti-phishing tool into a new product called SiteAdvisor. SiteAdvisor was available to end consumers and as part of their enterprise solutions. Like most industry offerings at the time, SiteAdvisor was a "lookup system" that relied primarily on blacklist databases of known phishing websites (e.g., PhishTank). In other words, if the link a user was about to visit appeared on a blacklist, a warning appeared. While this approach guaranteed low false positive warnings, reliance on list lookups was a reactive measure – many new/emerging threats didn't appear in the databases. Hence, true positive rates were also less than ideal. Some third party benchmarking studies conducted between 2007 and 2010 found that blacklist lookup systems had true positive rates under 80%, suggesting that one in five threats were slipping through the cracks. A VP for Product Innovation at McAfee at the time noted, "We have to build a better mousetrap."

As phishing threats persisted year over year, recognizing the need for proactive solutions, the entire industry began shifting towards hybrid solutions that coupled blacklists with machine learning models. Researchers at UVA had spent years developing better machine-learning models for anti-phishing. These models used thousands of "fraud cues" related to web page text, images, URLs, links, and source code to accurately predict whether a website was a phish with over 95% accuracy. Due to the cat-and-mouse adversarial nature of phishing, the anti-phishing feature sets and models had to be updated every hour to keep pace with new obfuscation strategies employed by fraudsters. In 2011, McAfee approached UVA to field test some of their anti-phishing machine learning models. By spring of 2012, a 3-month field test had been conducted. At the McAfee-UVA meeting to discuss the results, the VP for Product Innovation opened with the following:

> *"We have good news and bad news. The good news is that the phishing website detection model works exactly as advertised in our field studies with enterprise clients. The bad news is it didn't make a dent in terms of reducing user susceptibility rates…"*

Conventional wisdom had been that the reason users disregarded tool warnings was due to lack of trust in anti-phishing tools that simply weren't accurate enough. While this was true, and users did disregard the 95% accurate tool less than one that was only 80% accurate, disregard was no longer viewed as solely being a byproduct of poor performance. This notion was confirmed when subsequent experiments with a 99% accurate tool still caused users to disregard warnings 20% of the time. Pilot training programs educating users on phishing threats and anti-phishing tools helped in the short-term, but the effects wore off within three to four weeks. Completely blocking access to sites (i.e., not providing a continuation option when the anti-phishing tool deemed it a phish) resulted in customer churn in the consumer market and employee dissatisfaction in enterprise settings. In a few instances, the tool had blocked access to important work-related web pages.

By fall of 2012, there was a consensus that building a more accurate anti-phishing tool wasn't going to be enough, as highlighted by the VP for Product Innovation at the conclusion of one of the meetings:

> *"Clearly we need to go back to the drawing board. We need a multi-pronged approach that couples threat detection with some sort of intelligent user interventions. We know that warning disregard rates and overall user susceptibility follows the 80-20 rule, with 20% of users accounting for 80% of clicks on phishing links – this is true for employees and consumers. From conversations with our enterprise clients, we also know that these are some of their most productive and valuable employees. Given that security is always considered a secondary task, how can we find that sweet spot where we get them to stop clicking and transacting with so many darn phishing sites without annoying them or excessively disrupting their primary workflow?"*

Predicting Susceptibility to Phishing Attacks

Phishing susceptibility is when a user interacts with a phishing website. The question of *why* users are susceptible to phishing attacks has received attention from industry and academia. Folks at Carnegie Mellon University's CyLab developed the human-in-the-loop security framework (HITLSF) and the demographic-risk-knowledge model (DRKM). Some basic models had been developed in industry as well, including the ability-awareness model (AAM). However, these were largely explanatory models (descriptive analytics). It remained unclear whether susceptibility could actually be predicted. Given a user-phish encounter, could a machine learning model accurately predict whether the user would engage with the link (i.e., fall prey)? The UVA team received funding from industry and the National Science Foundation to study this exact question – whether phishing susceptibility prediction was possible, practical, and valuable.

After 12 months of research, the team felt that it was best to model phishing susceptibility in a manner analogous to e-commerce conversion funnels. Companies such as Amazon had become quite adept at predicting click-through rates and conversions for individual customers in real-time online session environments. Accordingly, the team developed a phishing funnel model (PFM) – depicted in Figure 1[1]. The model leveraged three types of features: tool, threat, and user-related. See Appendix A for detailed list of independent variables. It also featured four key target/dependent variables: visit, browse, consider legit, and intend to transact. Every time a user encountered a phishing link, PFM used a support vector ordinal regression model to predict the user's anticipated level of engagement with the phish on a 1-5 ordinal scale (no-visit, visit, browse, consider legit, transact), using objective and perceptual features. The perceptual features related to tool and threat perceptions and prior web experiences were captured via surveys. An initial model was trained on 908 university faculty, staff, and students using simulated e-commerce shopping tasks related to Internet banking and/or online pharmacies where each user was presented 10 search results – some legitimate, some phish. Users' funnel traversal behavior with respect to the phishing links were used to train the model.

[1] PFM combines the conversion funnel prediction task popularized by e-commerce firms, with the OODA loop concept developed by Colonel John Boyd of the USAF to understand decision-making in risky, real-time, adversarial settings.

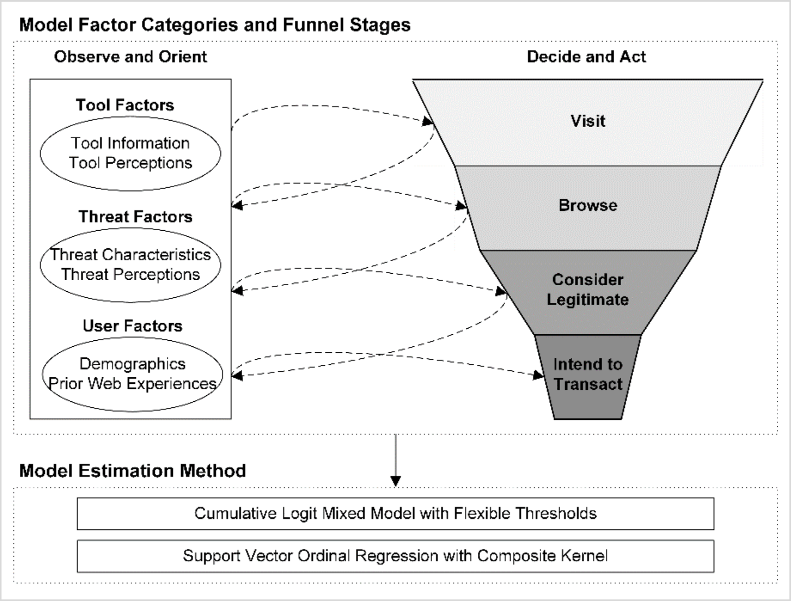


Figure 1: The Phishing Funnel Model (PFM)

Phase 1: Lab Experiments to Examine the Art of the Possible

As a first step, PFM was tested in a lab setting using 480 McAfee customers from their consumer segment. The model, trained in an academic lab setting, was examined on the McAfee customers. The experiment was conducted in the San Francisco area. PFM was compared against the aforementioned HITLSF, DRKM, and AAM susceptibility models which used different underlying feature sets. PFM was also compared against different classification methods including SVM, SVOR, CLMM, BayesNet, and LMM. Table 1 and Figure 2 present the AUC values from both analyses, which range from 0.5 to 1; higher values indicate better performance. PFM outperformed the three other susceptibility models (i.e., HITLSF, DRKM, and AAM) in terms of predictive power with an AUC 30% to 50% higher than its peers. Notably, the AUC values of DRKM and AAM were only marginally better than a random chance model. On the one hand, these results were promising: a model trained in an academic environment had performed well on McAfee customers (somewhat different demographics). On the other hand, the results were on a simulated Internet shopping task. It was unclear how well the model results would translate to enterprise field settings.

Table 1: AUC Values on Prediction ROC Curves for PFM and Comparison Models and Methods

Comparison Model	AUC
PFM	.847
CLMM-HITLSF	.608
CLMM-DRKM	.538
CLMM-AAM	.532

Comparison Method	AUC
PFM	.847
SVM	.728
SVOR	.719
CLMM-Equi	.710
BayesNet	.666
LMM	.620

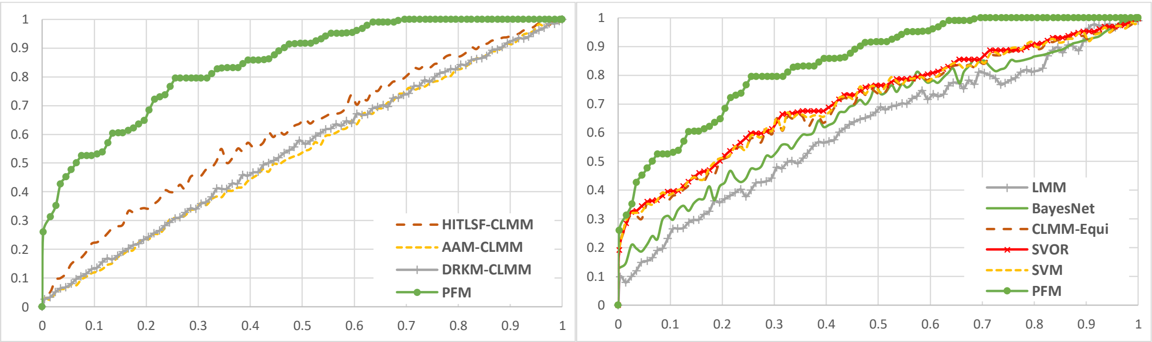


Figure 2: ROC Curves of Funnel Stage Predictions Across Models and Methods

Phase 2: Predicting "In the Wild" to Assess the Art of the Practical

In order to examine PFM's ability to predict users' phishing susceptibility over time and in organizational settings, a field experiment was conducted. The study was performed in two clients of McAfee Security: a large financial services company (FinOrg) and a mid-sized legal services firm (LegOrg). In each organization, employees with access to work-related computers were invited by high-level executives to participate in the experiment. Employees were not given details about the nature or purpose of the study – they were simply told that they would be asked to respond to quarterly surveys and periodically answer pop-up questions. In both companies, management incentivized employee participation by offering additional paid time off commensurate with participation duration. Table 2 provides an overview of the study participants; during the study's 12-month period, 50 participants (~4%) dropped out mostly due to normal turnover.

Table 2: Overview of Participants in Field Experiment

Company	Industry	Company Size	No. Invited	No. Participants	Opt-In Rate	Ave Age	Gender (Female)	Bachelor's Degree
FinOrg	Financial	Large	1151	796	69.2%	34.1	30.0%	90.1%
LegOrg	Legal	Mid-sized	655	482	73.6%	37.6	48.9%	86.5%
Total			**1806**	**1278**	**70.8%**	**35.4**	**37.2%**	**88.7%**

Each participant's Internet usage behavior on work-related computers—including encounters with potential phishing websites—was tracked using custom software. Figure 3 depicts the field experiment design. Surveys were administered longitudinally prior to the 1st, 4th, 7th, and 10th months and were used to gather participants' tool perception, threat perception, user experiences, and demographic information from PFM, plus all items related to HITLSF, DRKM, and AAM, as done in the lab experiment. Unlike the previous experiment in which the domains were limited to banks and pharmacies, the field experiment encompassed a broader array of domains; moreover, for all phishing URLs encountered by participants during the field experiment, threat category variables including threat domain, threat type, and threat severity were derived automatically.

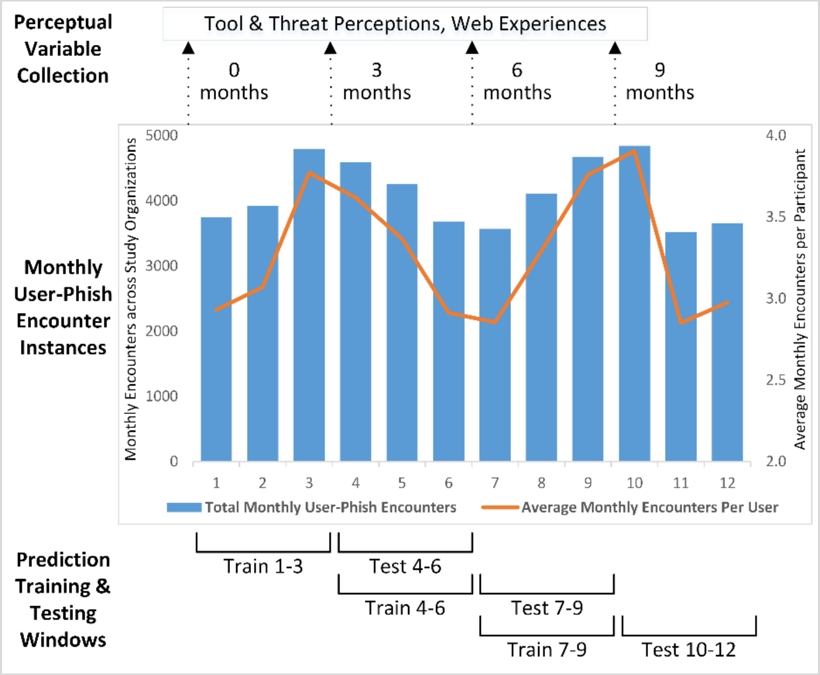


Figure 3: Illustration of Field Experiment Design

During the field experiment, all of FinOrg's participants' work computers were equipped with an enterprise endpoint security solution capable of detecting email and web-based phishing threats using robust rule-based and machine learning-driven analysis of URLs and website content. This solution leveraged both client-side servers coupled with McAfee's machine-learning servers. Similarly, for the duration of the field experiment, LegOrg partcipants' work computers were equipped with an endpoint protection solution designed for small-to-medium-sized businesses. This was a lighter solution without the need for constant interaction with McAfee's servers. The detection rates and run times for the FinOrg tool where 96% and 900 milliseconds, respectively, whereas the LegOrg anti-phishing tool performed at 87% and 1.9 seconds. Both software packages displayed prominent warnings whenever a URL deemed to be a potential phish was clicked.

Since the field experiment occurred in real-time as participants interacted with websites on their work computers, a mechanism was necessary to collect funnel stage variables from all *potential* phishing websites irrespective of whether the website had been verified as phishing or not. Consequently, for the field experiment, a URL was operationalized as potential phishing if (a) the organizations' endpoint security tool considered it to be a phish, in which case a warning would appear; or (b) the URL appeared in any of several reputable phishing website databases as either verified or pending based on a real-time check. Funnel stages were also determined for each potential phishing URL in a manner consistent with the controlled experiment. Visitation and browsing decisions were automatically recorded from clickstream logs. Moreover, once a participant had concluded their session with a potential phishing site, a pop-up form asked them if they considered the site legitimate and/or intended to transact with the site.

All potential phishing URLs encountered during the 12-month period were eventually verified against online databases, resulting in a test bed of verified participant-phish encounters. The 1278 participants ultimately encountered 49,373 total verified phishing URLs during the 12-month period during which complete funnel stage behavior data was captured. As depicted using the bar and line chart in Figure 3, this averaged out to ~3.25 URLs per participant per month (or 4,100 mean monthly participant-phish encounter instances). Due to the longitudinal nature of the participant-phish encounters, a windowing approach was used for model training and testing (see the bottom of Figure 3). Initially, all instances from the first three months including the survey results at 0 months were used to train PFM and the comparison models. These models were applied to the participant-phish encounters that transpired during months 4-6 (using survey data from "3 months" [top of Figure 3] for tool/threat perceptions and web experiences). Next, the 4-6 month data were used as training data for a month 7-9 test window, and so on.

PFM was again compared against the HITLSF, DRKM, and AAM susceptibility models, along with the SVM, SVOR, CLMM, BayesNet, and LMM methods. All models and methods were evaluated on the 36,909 test instances that transpired in the last nine months (i.e., months 4-12). As shown in Table 3 and Figure 4, PFM outperformed the three comparison models with AUC values that were 22% to 35% higher, and PFM's AUC was also between 8% and 25% higher than the competing susceptibility prediction methods.

Table 3: AUC Values on Prediction ROC Curves for PFM and Comparison Models and Methods

Comparison Model	AUC
PFM	.875
CLMM-Flex-HITLSF	.642
CLMM-Flex-DRKM	.562
CLMM-Flex-AAM	.548

Comparison Method	AUC
PFM	.875
SVM	.761
SVOR	.753
CLMM-Equi	.730
BayesNet	.681
LMM	.629

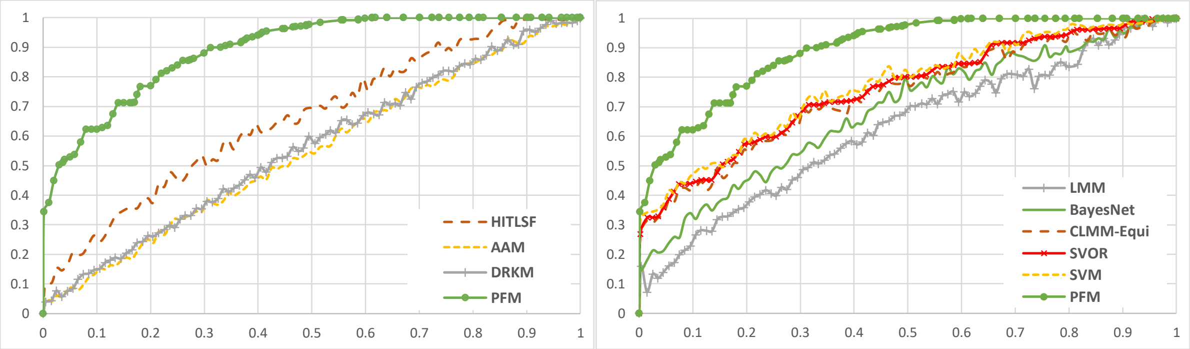


Figure 4: ROC Curves of Funnel Stage Predictions Across Models and Methods

The team analyzed the detection performance of PFM and the top-performing comparison model (HITLSF) and method (SVM) using the 1,421 intention to transact instances that transpired during the 9-month test period. The left bars in Figure 5 depict the number and percentage of correctly classified intend-to-transact instances, with PFM detecting 10% to 17% more instances than its best competitors. They also extracted a subset of these instances where some transaction behavior was "observed" via the log files, amounting to 1,165 transactions where the employee either entered information (e.g., in a form or login text box) or agreed to download files or software to the work machine. These observed transactions were examined to see how many were predicted as intention (i.e., the most severe stage in the funnel). As shown in the right bars in Figure 5, PFM also attained markedly better performance on this subset of observed transactions, with detection rates of 90 to 94%.

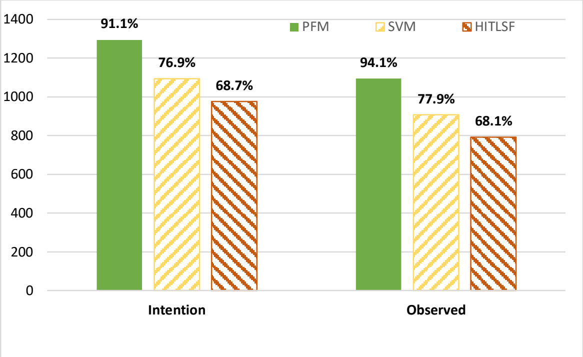


Figure 5: Number and Percentage of Correctly Predicted Employee Intention to Transact Instances

Regarding visits to high severity phishing URLs containing malware, Figure 6 depicts the frequency of concocted (Con) and spoof (Spf) sites where PFM, SVM, and HITLSF correctly predicted that the user would at least visit the URL. The bars denote threats encountered via email (work or personal), social media, or search engine results, and threats were also categorized as generic attacks (Gen), spear phishing attacks (SP) tailored towards the organizational context, or wateringhole attacks (WH) that leverage concocted websites. As depicted, PFM outperformed the best comparison model (HITLSF) and method (SVM) on high-severity threats across various communication channels, with the exception of generic spoof attacks appearing in work email. Overall, PFM was able to correctly predict visits to high severity threats for 96% of the cases in the 9-month test period, which amounts to 170 greater detection occurrences (10% points higher) than the closest competitor. Analysis of performance within these different threat channels revealed that PFM was fairly robust across email, social media, and search engine threats.

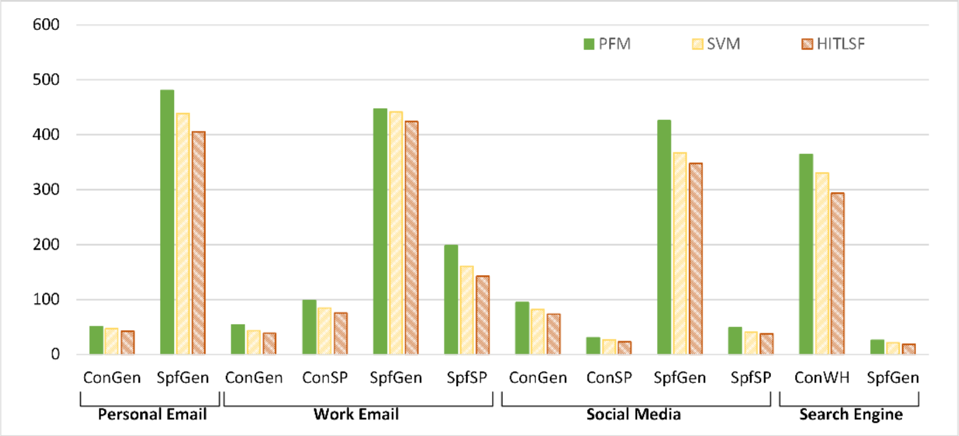


Figure 6: Number of Correctly Predicted High Severity Threats Visited by Employees

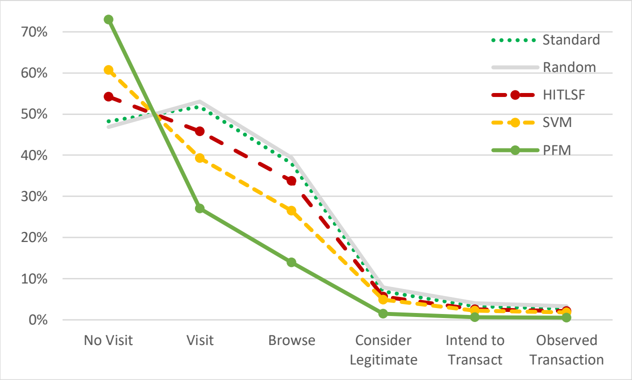
Phase 3: Interventions to Scratch the Art of the Valuable

While the 12-month prediction field study demonstrated the predictive potential, a critical question remained unanswered: “How effectively can interventions driven by susceptibility predictions improve avoidance outcomes?” To answer this question related to the downstream value proposition of accurately predicting susceptibility, the team followed up the prediction field experiment with a longitudinal multivariate field intervention experiment. The field test was performed over a three-month time period at FinOrg and LegOrg, using the same set of employees incorporated in the prior field experiment. Due to normal workforce attrition and a few opt-out cases, 1,218 employees participated in the experiment. The experiment design and variable operationalizations utilized were the same as the prior field study. All participants filled out the same survey as prior experiments at the beginning of the 3-month period.

Each participant was randomly assigned to one of five settings for the duration of the experiment: *PFM, SVM, HITLSF, Random*, and *Standard*. Employees in the *Standard* setting represented the status quo control group: these individuals received the default McAfee warning for each phishing URL, irrespective of their predicted susceptibility levels. Conversely, the *PFM*, *SVM*, and *HITLSF* groups received one of three warnings (default, medium, and high severity) based on their respective model’s predicted susceptibility level along the phishing funnel. Aligning warnings with user or other contextual factors has been found to be a potentially effective security intervention, provided warning fatigue can be properly managed. These warnings differed in terms of size, colors, icons, and severity of message text. For user-phish encounters predicted to end without a visit, the default McAfee warning was displayed. For those predicted to result in visitation and/or browsing, the medium severity warning was presented. Finally, user-phish encounters predicted to culminate with consider legitimate or intend to transact garnered a high severity warning. In order to control for behavioral changes attributable to introduction of the new medium and high severity warnings, relative to the default one used in the *Standard* setting, an additional *Random* setting was incorporated. Participants assigned to this setting randomly received either the default, medium, or high severity warning. For those employees assigned to the *PFM*, *SVM*, and *HITLSF* settings, data from months 10-12 of the prior experiment was used to train their respective susceptibility prediction model. To reiterate, model predictions were not used for employees in the *Random* and *Standard* settings. During the three-month study, there were an average of 11.35 actual phishing encounters per employee.

Performance was evaluated by examining actual phishing funnels for participants assigned to the five settings. Figure 7 depicts the experiment results. Participants using *PFM* for susceptibility prediction were less likely to traverse the phishing funnel stages, with lower visitation, browsing, legitimacy consideration, and transaction intention rates. On average, *PFM* outperformed *SVM, HITLSF*, and

Standard by 7 to 20 percentage points on the higher funnel stages and garnered less than half the number of traversals for the latter stages of the funnel. The users assigned to the benchmark or baseline settings had 3 to 6 times as many observed transactions with phishing websites across the 3-month duration of the study, relative to users assigned to PFM. These results highlight the sensitivity of intervention effectiveness to the performance of the underlying predictive models' accuracy in field settings, thereby underscoring the importance of enhanced prediction. Interestingly, the *Random* setting underperformed the *Standard* setting, suggesting that displaying alternative warnings without aligning them with predicted susceptibility levels did not improve threat avoidance performance. The results highlight the downstream implications and efficacy of accurately predicting phishing susceptibility and suggest that such predictive analytics can be a viable component of an enterprise anti-phishing strategy.



Funnel Stage	PFM	SVM	HITLSF	Random	Standard
Visit	27.02	39.27	45.78	53.08	51.76
Browse	13.88	26.46	33.72	39.37	38.05
Consider Legitimate	1.45	4.83	5.59	7.84	6.92
Intend to Transact	0.61	2.24	2.50	4.03	3.25
Observed Transaction	0.50	1.74	2.10	3.27	2.64

Figure 7: Phishing Funnel Percentages for Employees Assigned to Five Experiment Settings

The Path Forward

Collectively, the lab, field, and multivariate experiments demonstrated the potential for predicting susceptibility. In order to offer this as a product in enterprise settings, folks at McAfee knew they needed to conduct a cost-benefit analysis to better quantify the value proposition for their clients, inform possible pricing options, and illuminate their go-to-market strategy. It had been a long journey, and much work remained to be done, but there *appeared* to be light at the end of the tunnel.

As a next step, they decided to examine a visit/no visit binary prediction model's annual cost-benefit for FinOrg. The model would display a higher severity warning to employees predicted to visit a particular phishing URL. For your analysis, you may use the following assumptions:

- FinOrg has 30,000 employees
- On average, each employee encounters 3 phishing URLs per month
- FinOrg believes that displaying a high severity warning unnecessarily (i.e., when the user was not going to visit the URL) will reduce productivity by 1 hour due to employees feeling annoyed, seeking HelpDesk support, clarification, etc. Such incidents cost the firm an estimated $50, each. Although employee attrition over time is an annoyance byproduct, you may assume it is not critical for this analysis.
- Approximately 1% of those visiting a phishing URL will eventually transact (i.e., provide information, download malware, etc.). Each such transaction costs FinOrg $30,000 on average.
- Although the Phase 3 experiment revealed a 0.5% transact rate, for the purpose of your cost-benefit analysis, you may assume that displaying a high severity warning to someone that *was* planning to visit the phish will result in complete success – 0% visitation rate.

Exhibit 1

Exploring a PhishCasting Capability at McAfee

Table A1: Features (Independent Variables) used in PFM

Category	**Sub-category**	**Variables**	**Description**
Tool Factors	Tool Information	Tool Warning	Whether or not the anti-phishing tool displayed a warning
		Tool Detection Rate	The accuracy of the anti-phishing tool
		Tool Run Time	The time, in seconds, needed by the machine-learning-based tool to make a prediction regarding whether a given URL is a phish or not
	Tool Perceptions	Tool Usefulness	Survey-based items related to user perceptions regarding the usefulness of their anti-phishing tool
		Trust in the Tool	User's level of trust in the anti-phishing tool
		Tool Effort Required	Survey-based items related to user perceptions regarding the level of effort needed to use the anti-phishing tool
		Cost of Tool Error	Survey-based items related to user perceptions regarding the cost of false positives/negatives of their tool
Threat Factors	Threat Characteristics	Threat Domain	The URL domain (e.g., financial services, e-commerce, social media, etc.)
		Threat Type	The type of phishing attack, such as spoof, concocted, etc.
		Threat Severity	The level of severity of the phishing URL (e.g., identity theft, malware, etc.)
		Threat Context	Where the threat appears, such as email, search result, social media, etc. Here we focus on position in display (e.g., "7th unread email in inbox" or "4th search result")
	Threat Perceptions	Phishing Awareness	Survey-based items related to user perceptions regarding their level of awareness of phishing threats
		Perceived Phishing Susceptibility	Survey-based items related to user perceptions regarding how susceptible they consider themselves to phishing
		Perceived Phishing Severity	Survey-based items related to user perceptions regarding how severe they consider phishing threats to be, in general
User Factors	Demographics	Gender	Gender of the user
		Age	Age of the user
		Education	Education level of the user
	Prior Web Experiences	Trust in Institution	Survey-based items related to user perceptions regarding their level of trust of relevant institutions such as banks, pharmacies, etc.
		Trust in Web	Survey-based items related to user perceptions regarding their level of trust in the Internet
		Familiarity with Domain	Survey-based items related to user perceptions regarding their level of trust in the websites' domain (e.g., financial services, e-commerce, etc.)
		Familiarity with Site	Survey-based items related to user perceptions regarding their level of familiarity with the site (e.g., Bank of America's website)
		Web Activities	Summative score of web activities such as social media, online shopping, blogging, forums, etc.
		Security Habit	A score of user's security habits based on observed logs
		Self-Efficacy	Survey-based items related to belief in one's abilities
		Risk Propensity	Survey-based items related to user's risk propensity
		Past Encounters	Self-reported past encounters attributable to phishing attacks
		Past Losses	Self-reported prior losses attributable to phishing attacks