Exposing Decision Biases Committed by Machine Learning Algorithms: Revisiting the Boy Who Cried Wolf in the Context of Phishing Detection

C.J.Duan
Dulun Consulting Group, research@dulun.com, http://www.dulun.com

Grown out of the quest for artificial intelligence (AI), machine learning (ML) is today's most active field across disciplines with sharp increase in applications ranging from criminology to fraud detection and to biometrics. Machine learning and statistics both emphasize the importance of model estimation / training and thus share the inescapable type I and II errors. Extending the concepts of statistical errors into the domain of machine learning. we devise aground-breaking pH scale (in chemistry)-like Litmus ratio and intend it as a litmus test of decision bias masked by the established criterion of Accuracy. Using publicly available phishing data set, we conduct experiments on a series of single-feature models using the CHAID package in R. Based on the results, we recommend practitioners match the risk over / under estimation cost ratio with the error ratio associated with each machine learning model in order to mitigate potential losses in their specific decision-making environments

Key words: machine learning, Type 1 and 2 errors, CHAID, Litmus ratio, decision bias, phishing websites classification
History: submitted on April 4, 2018

1. Introduction

Aesop's fable, the The Boy who Cried Wolf, is about a young shepherd boy found his life dull in the pasture as he sat on the hillside tending his masters sheep. To amuse himself, he ran toward the village shouting at the top of his voice, ”Wolf! Wolf! The Wolf is chasing the sheep!”. The villagers dropped their work and rushed up the hill to help the boy scare the wolf away. But upon their arrival, they found no wolf and only the grinning face of the shepherd boy. ”Don't cry 'wolf', when there's no wolf!” The grumbling villagers told the boy and returned to their village. A few days later, the boy felt bored and yell out again, ”Wolf! Wolf! The wolf is chasing my sheep!” To his wicked delight, he viewed the villagers dash up the hill and fall for his trick gain. The villagers sternly warned ”Don't cry 'wolf' when there is NO wolf!”. Later, he saw a real wolf prowling about his herd. In terror, he leaped to his feet and sang out as loudly as he could, ”Wolf! Wolf!”. To no avail, the villagers thought he was trying to repeat his foolish game, so they stay away. At sunset, everyone wondered why the shepherd boy hadn't returned to the village with their sheep. They went up the hill and found him weeping. The price the village paid is costly: the killing of a great many of the flock.

Table 1 The Villagers Benefit and Cost Matrix

	Perception	
Reality	Wolf - Yes	Wolf - No
Wolf -Yes	Respond and Rescue 6 villagers for 1 hr @ \$10 h^{-1}	Stay Put Killing of 3 sheep @\$200 each
Wolf -No	Respond and Disappointed 6 villagers for 1 hr @ \$10 h^{-1}	No Action and No Harm \$ 0

Adapted from Roulston and Smith (2004)

At the end of the story, an old man comforted the boy. If you tell too many lies, no one believes you when you tell the truth. The moral message the tale conveys is that liars are not trustworthy even when they are telling the truth. Undoubtedly, the shepherd boy was not even close to a pure rational decision maker, perhaps a naughty one. Considering his maturity level and working environment, are we being too harsh on the young boy entrusted with the task of guarding the villages most valuable asset? Roulston and Smith (2004) delineate the classical tale in a decision-making matrix as shown in table 1. Based on their cost-loss analysis, the villagers in the tale were unprepared to tolerate a reasonably high false-alarm rate (cries of Wolf! when there is no wolf looming). As revealed by our ensuing investigation, even machine learning (ML) models trained on large data sets are not immune from sounding false alarms.

Our research was motivated by what we observed as a slight disconnect between the computing side (predictive accuracy) and statistical side (model bias) of ML. ML should induce more objective and evidence-based decision-making, since machines are supposedly free from human prejudice. However, as revealed in the work of Caliskan et al. (2017), machine learning (language processing) can acquire stereotyped biases from textual data reflecting everyday human culture. As noted by Kleinberg et al. (2017), applied ML work typically focuses on the link between data and prediction, while paying less attention to the prediction→decision link. The limitations of any model trained on real data necessitates human experts' judgment back into the decision-making processes - for the worse or the better (Lewis 2016).

Purported to fill this gap, our study was undertaken with four overall goals:

1. To alert the decision analysis community to the prevalence and peril of model bias in ML.
2. To propose and formulate an Litmus ratio- ι (akin to the pH scale in chemistry) for the assessment of decision bias committed by ML models.
3. To demonstrate empirically model bias is unrelated to accuracy and requires human judgment and stringent regulation.
4. To illustrate how human adopters of ML models can use our proposed Litmus ratio to diagnose and mitigate bias implicit in ML.

The paper proceeds as follows. In the Related Works section, we describe phishing detection, which serves as the context of our experiments. Also, discussed in that same section are machine learning and statistics, statistical errors, and Chi-squared Automatic Interaction Detection (CHAID) - the machine learning algorithm we used for our project. Amid the Related Works section, we proposed a unique alpheta ratio for the purpose of assessing machine learning models. In the Methodology section, we describe the Phishing data set, the set of 12 models to be evaluated, as well as the K-fold cross validation used to assess models. The Results section presents the results we acquired from out experiments. In the Implications section, we discuss how to match a machine learning models alpheta ratio to the risk overestimation cost and underestimation loss profile dictated by the decision-making environment. In the Concluding Remarks section, we reexamine the Boy Who Cried Wolf fable via alternative lens of decision making and spotlight the importance of human intelligence in the face of wide-spreading artificial intelligence.

2. Related Works

2.1. Classification using ML Algorithms

Mitchell (1997) defined machine learning as, "a computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance in task T, as measured by P, improves with experience E" (p. 2). Machine learning is the manifestation of statistical learning (statistics) algorithms implemented via software (computing) applications. The concept of machine learning has been in existence for decades. But it was only until recently that machine learning began to see its explosive applications to huge quantities of data attributed to the confluence of powerful computers, cheap storage of data, and vast availability of analytical opportunities, among others. Being a subfield of artificial intelligence, machine learning rests on the paradigm that algorithms can learn from data and reason with data (Rao and Govindaraju 2013). Based on the desired outcome and the type of input available for training the system, machine learning algorithms generally fall into either supervised or unsupervised learning. In the former, a trainer carefully selects examples for the learner and the learner must sort those examples into categories specified by the trainer, whereas in the latter the learner is given little or no instruction on the learning task and the goal is to find underlying regularities (Cottrell 2006).

There exists considerable overlap between statistics and machine learning. Both fields focus on studying generalization from data (model building). However, statistics, inferential statistics in particular, makes generalization (prediction) about the unknown population parameters (parameter estimation) based on often small sample. Machine learning, on the other hand, relies heavily on the predictive accuracy of models, while ignoring largely checking of models and assumptions. In addition, terminology employed in machine learning is different from that used by statisticians.

In machine learning, a target or outcome is called a label, while in statistics it is referred to as dependent variable (DV). In statistics, an input predictor is called an independent variable, whereas it is denoted as a feature in machine learning.

In a classification ML problem, the focal algorithm is presented a training data sample of n observations on a categorical (class) variable (label) Y and a set of m predictor variables X $\{X_1, X_2, \cdots, X_m\}$ (features) that are either categorical or continuous. The goal of the learning process is then to find a model predicting Y values from those of X (Loh 2011). Classification tree methods typically partition recursively the training dataset one X_i at a time to achieve minimal impurity (measure of misclassification).

Originally proposed by Kass (1980), the acronym CHAID stands for Chi-squared Automated Interaction Detection. It aims to detect interactions between categorized variables of a data set, one of which is the outcome variable or label. CHAID is a classification tree algorithm to construct a non-binary decision tree, where two or more branches can attach to a single node. Because the CHAID algorithm will often effectively build many multi-way frequency tables (e.g., when classifying a categorical response variable with many classes, based on categorical predictors with many classes), small sample size may not yield optimum prediction. The algorithm is implemented based on recursive partitioning and is purported to maximize the significance of a chi-squared statistic for cross-tabulations between the dependent variable and the predictors at each partition. The data are partitioned into mutually exclusive, exhaustive subsets that best describe the outcome variable. Multiway splits are used by default. It is non-parametric and quite proficient in handling large dataset.

2.2. Phishing Detection

Phishing is defined as a social engineering technique where the aggressor pretends to be a trustworthy entity in attempt to gain sensitive information (e.g. usernames, password, credit card credentials) from the victim (Jagatic et al. 2007). In the cyber-attack, fraudulent website con users into disclosing sensitive and/or confidential information to the attacker by impersonating to be a legitimate website in an automated fashion. These types of communications are commonly accomplished through emails that point users to phishing websites through which phishers gather information in question. Bank and credit card details, password, and personal identification number are few examples that interest phishers frequently collect users credentials (Jakobsson and Myers 2006)

Phishing websites has been a persistent threat and it has been a major concern of cyber security. According to the report published by Anti Phishing Working Group (APWG) in December 2016, the unique phishing sites detected only in 3rd quarter of 2016 are 364,424. The most common methods of phishing involve some forms of technical deception designed to make a link in an email (and the spoofed website it leads to) appear to belong to the spoofed organization.

In reviewing the applied ML research on phishing websites classification, link features (presence of special symbols, and URL length, etc.).

2.3. Evaluation of classification ML Algorithms

In assessment of (binary) classification ML algorithm performance, the predicted values and actual label values are often cross-tabulated into a typical confusion matrix (CM) contained in table 2. Our acquaintance with null hypothesis significance testing (NHST) easily alludes us to Type 1 (α) and Type 2 error (β). Type 1 error, also known as false positive, is the false rejection of a null hypothesis when it is true. Plainly speaking, it arises when we are detecting a difference when there is none. Type 2 error, also known as ”false negative”, is the error of not rejecting a null hypothesis when the null hypothesis is the false state of reality. In other words, it occurs when we are failing to detect a difference when the difference in fact exists. As described below, all common metrics in the assessment of ML classifying models are derived from such CM.

Accuracy is the most intuitive performance measure of a ML classification model and it is simply the ratio of correctly predicted observation (both phishing and legitimate) to the total observations in the test dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Precision, also known as PPV (Positive Predictive Value) is the rate of correctly predicted positive labels to the total predicted positive observations. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as Sensitivity or True Positive Rate (TPR), is is the ratio of correctly predicted positive observations to the all observations in actual “+” class:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Specificity, also known as True Negative Rate (TNR), is the ratio of correctly predicted “-” observations to the all observations in actual“-” class:

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

F1 score is the weighted average or harmonic mean of Precision (P) and Recall (R). Thus, it takes both false positives and false negatives into account, as shown in equation (5).

$$F1 = \frac{2 * P * R}{P + R} \tag{5}$$

Table 2 Sample Confusion Matrix

	Predicted Value	
Actual	+	-
+	True Positive (TP)	False Negative (FN)
-	False Positive (FP)	True Negative (TN)

Adapted from Thabtah et al. (2016)

After intensive review of metrics deployed in assessment of ML classifiers, we found no metrics involve only FP and FN in definition. Inspired by this obvious absence of bias measuring metrics, we, in the next section, take the first step and attempt to formulate a simple Litmus ratio that can be applied in evaluating ML model bias.

3. The Proposed Litmus Ratio

We now formally define the Litmus ratio $\tau = \frac{\alpha}{\beta}$, where α is the total number of false positive cases misclassified and β is the false negative counterpart within the test data set of size N. If $\mathit{\Lambda}$ is the accuracy of the corresponding ML model, we can then establish

$$1 - \mathit{\Lambda} = \frac{\alpha + \beta}{N} \tag{6}$$

Substitute α with $\tau\beta$, we arrive at:

$$1 - \mathit{\Lambda} = \beta\frac{1+\tau}{N} \Rightarrow \beta = N * \frac{1-\mathit{\Lambda}}{1+\tau} \tag{7}$$

With C_{FP} and C_{FN} represent unit cost (negative) of FP and FN error respectively, we can easily calculate the total cost of misclassifications ($\mathit{\Gamma}$) as:

$$\mathit{\Gamma} = N * \frac{(1-\mathit{\Lambda})\tau}{(1+\tau)} * C_{FP} + N * \frac{1-\mathit{\Lambda}}{(1+\tau)} * C_{FN} \tag{8}$$

For simplicity of presentation, we use $k = \frac{C_{FP}}{C_{FN}}$ and further reduce equation 3 to :

$$\mathit{\Gamma} = \alpha * C_{FP} + \beta * C_{FN} = N * (1-\mathit{\Lambda}) * C_{FN} * \frac{(k\tau + 1)}{(1+\tau)} \tag{9}$$

Consider a special case of $k = 1$ (neutral decision-making environment), we have $\mathit{\Gamma} = N * (1 - \mathit{\Lambda}) * C_{FN}$, which indicates that the best model is the one with the highest accuracy only and bias is canceled out.

Without any further calculus work, we can see the close ties between model characteristics and the cost characteristics of the corresponding decision-making environment, represented by $\tau(\tau'), \mathit{\Lambda}$ and k respectively. The purpose of next section is to empirically demonstrate the $(\mathit{\Lambda} - \tau)$ - k links, which can be applied to the screening of ML model candidates.

Table 3 **Performance Results**

Model	Features	Λ	τ
1	01 - 12	90%	0.78
2	13 - 18	88%	0.24
3	19 - 23	57%	0.01
4	24 - 30	73%	0.57
5	01 - 30	95%	1.27

Models differ in features included

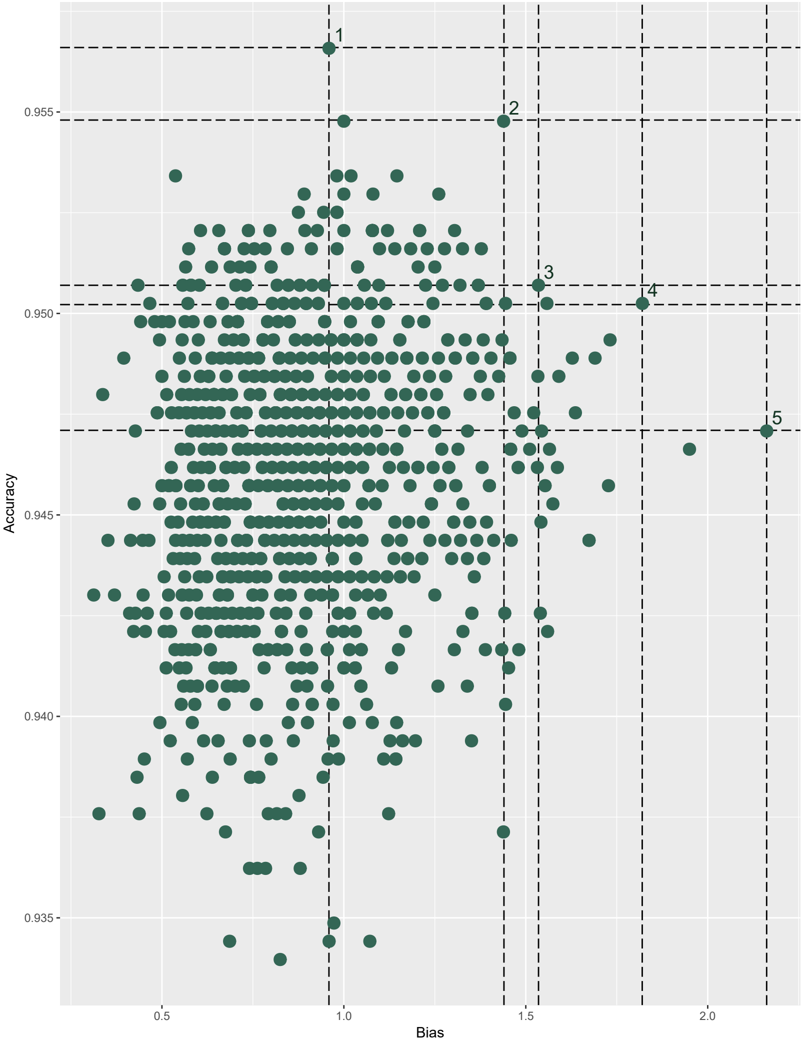
4. Data, Models and Empirical Analyses

The Phishing Websites Dataset (Mohammad 2015) , used in this work, is available online at the UCI Repository (Dheeru and Karra Taniskidou 2017). The dataset has been used in several works that either propose new or evaluate existing techniques for phishing website classification (Thabtah et al. 2016, Mohammad et al. 2012, 2013, 2015). It contains data from 11055 websites labeled into two major classes (-1 for phishing and 1 for legitimate). Of the 11055 observations, 44% take the value of -1 (phishing) and 56% take the value of 1 (legitimate). Each instance contains 30 features / attributes, which fall into four broad categories. Feature 1 through 12 belong to the Address Bar Based Features category Abnormal Based Features occupy the next 6 features. Feature 19 - 23 are bundled into the HTML and JavaScript Based Features, while Domain Based Features claim the remaining 7 attributes. All 30 features have been preprocessed into either dichotomous (-1 for phishing positive, 1 for phishing negative) or trichotomous (-1, 1, and 0 for phishing neutral).

Customarily, we split the whole dataset into 80% for training and 20% (size of 2211) for testing. We carried out experimental analyses on the CRAN-R platform (R Core Team 2013). In the first phase, we feed the whole training dataset of size 8844 into the CHAID algorithm implemented by the R-Forge Project Team (R-Forge 2015), while restricting the variety of features involved in model training.

We summarized the performance of each of the five models assessed on the test dataset in table 3. The full model with all 30 features outperformed the other 4 sub-models by the margins ranging from 5% to 38% in terms of accuracy. Model 3, which considers only the HTML and JavaScript Based Features, yielded the lowest accuracy and extremest bias. Like regression in statistical analysis, we observe that individual (categories of) features do not contribute equally to model accuracy. It is also noteworthy that the full model (model 5) not only improves accuracy to the highest level, but also shifts the bias from the < 1 side to the > 1 side. It seemingly confirms that in ML, the more objective the data and the larger the data set, the less possibility of distortion, as noted by Rosso (2015).

Figure 1 Bias - Accuracy Scatter Plot



Note. each dot represents one model with unique parameterization

In the second phase of our empirical analyses, we created 1,000 models, which consider the same set of 30 available features, but differs mainly in the records contained in the training data. Via iterative re-sampling, we created 1,000 different samples of size of 7075 (80% of 8844) from the original training dataset and fed each sample into the CHAID algorithm. The 1,000 models are plotted in Figure 1 along the dimensions of accuracy and bias. This arrangement allows us to investigate the impact of various accuracy-bias combinations on the total cost of misclassification (Γ).

The context of phishing detection resembles the situation faced by the villagers in the opening fable, in that C_{FN} far outweighs C_{FP}. For illustration purposes, we set k = 1/10 and computed the Γ for each of the five models in the top-right quadrant. Surprisingly, we noticed in Table 4 that the best model with the lowest Γ of 44.49 is model #5, which has the worst accuracy of 94.7%

Table 4 **Accuracy and Bias Tradeoff**

Model	N	Λ	k	τ	Γ
1	2211	0.957	1/10	0.96	53.66
2	2211	0.955	1/10	1.44	46.86
3	2211	0.951	1/10	1.54	49.60
4	2211	0.950	1/10	1.82	46.13
5	2211	0.947	1/10	2.16	44.49

All models with the same set of features
Total Costs are in units of $C_{FN}, k = \frac{C_{FP}}{C_{FN}}$

among the five models studied. The model with the highest accuracy of 95.7% (model #1) yielded the worst total cost figure of 53.66. The worsened predictive accuracy is well compensated by the right-shifting bias factor.

5. Implications

Because we utilized the same hold-out dataset to test all models, we are afforded the opportunity to unravel the familiar type I and II errors in statistics in the context of machine learning. Completely masked by the overall misclassification rate, type I and II errors take on significantly different meanings in the evaluation of machine learning models. Regarding labeling phishing websites, a type I error means a legitimate website is wrongly put on blacklist. Consequently, this overestimation of the potential risk posed by the candidate website will induce wasted resources, most of which are largely opportunity costs (the extra hour spent finding alternative websites to get the original job done). On the other hand, underestimation of potential phishing threat will definitely deal identity theft and resultant financial loss and psychological stress to the victim. If we assume an individuals identity and fanatical safety is one of the most critical assets and the extra time wasted on searching for alternative websites due to false alarms can be reasonably expressed in hourly wages, the Overestimation Cost (OC) -Underestimation Loss (UL) analysis should give a numerical ratio value as calculated in ....

6. Concluding Remarks

George Box, “one of the great statistical minds of the 20th century”, made the famous quote - “all models are wrong, but some are useful” (Box and Draper 1987, p. 424). The aphorism today still shines a somewhat colder light on the much hyped and feared might of machine learning and AI. Decision analysis is based on the maximum expected utility (MEU) action axiom, the basic proposition of which states that human beings purposely utilize means in order to pursue desired ends. To make high quality decisions, a decision-making agent (partnership between man and machine) must have a sense of the association between various actions and the likelihood of different outcomes, as well as the desirability of each such outcome. Correspondingly, they represent prediction (from input features to output labels) and judgment (assign values to the expected utility of models derived). On the one hand, latest advances in AI vastly reduced the cost of prediction. On the other hand, AI also substantially raised the value of human judgment. To this end, we put forward our own version of Box's apothegm - "All models are biased, some are aggreably preferreed"

Acknowledgments

References

Box GE, Draper NR (1987) *Empirical model-building and response surfaces.* (John Wiley & Sons).

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186, URL `http://dx.doi.org/10.1126/science.aal4230`.

Cottrell GW (2006) New life for neural networks. *networks* 5:6.

Dheeru D, Karra Taniskidou E (2017) UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

Jagatic TN, Johnson NA, Jakobsson M, Menczer F (2007) Social phishing 50:94–100, ISSN 0001-0782, URL `http://dx.doi.org/10.1145/1290958.1290968`.

Jakobsson M, Myers S (2006) *Phishing and countermeasures: understanding the increasing problem of electronic identity theft* (John Wiley & Sons).

Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. *Applied statistics* 119–127.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. *The Quarterly Journal of Economics* URL `http://dx.doi.org/10.1093/qje/qjx032`.

Lewis M (2016) *The undoing project: A friendship that changed our minds* (WW Norton & Company).

Loh WY (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(1):14–23, URL `http://dx.doi.org/10.1002/widm.8`.

Mitchell TM (1997) Machine learning.

Mohammad RM (2015) Phishing websites features. URL `http://dx.doi.org/10.13140/rg.2.1.2595.6000`.

Mohammad RM, Thabtah F, McCluskey L (2012) An assessment of features related to phishing websites using an automated technique. *Proc. Int. Conf. for Internet Technology and Secured Transactions*, 492–497.

Mohammad RM, Thabtah F, McCluskey L (2013) Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications* 25(2):443–458, URL `http://dx.doi.org/10.1007/s00521-013-1490-z`.

Mohammad RM, Thabtah F, McCluskey L (2015) Tutorial and critical analysis of phishing websites methods. *Computer Science Review* 17:1–24, URL `http://dx.doi.org/10.1016/j.cosrev.2015.04.001`.

R Core Team (2013) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

R-Forge P (2015) *CHAID: CHi-squared Automated Interaction Detection.* R package version 0.1-2.

Rao CR, Govindaraju V (2013) *Handbook of Statistics, Volume 31: Machine Learning Theory and Applications* (North Holland & IFIP).

Rosso C (2015) The conundrum of machine learning and cognitive biases. *Medium* .

Roulston MS, Smith LA (2004) The boy who cried wolf revisited: The impact of false alarm intolerance on costloss scenarios. *Weather and Forecasting* 19(2):391–397, URL http://dx.doi.org/10.1175/1520-0434(2004)019<0391:TBWCWR>2.0.CO;2.

Thabtah F, Mohammad RM, McCluskey L (2016) A dynamic self-structuring neural network model to combat phishing. *2016 International Joint Conference on Neural Networks (IJCNN)*, 4221–4226, URL http://dx.doi.org/10.1109/IJCNN.2016.7727750.