Submitted to *Decision Analysis*
manuscript

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

C.J.Duan
DulunConsulting Group, research@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 I and II errors, CHAID, Litmus ratio, decision bias, CRAN-R, phishing websites classification
History:

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 Time and Efforts well spent	Stay Put Loss of Assets
Wolf -No	Opportunity Cost Waste of Time and Efforts	No Harm Zero Cost

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. 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 discuss how human model adopters can use our proposed Litmus ratio and tailor ML models to the circumstances

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 re-examine 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.

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)

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 ML Algorithms

In assessment of ML algorithm performance, the predicted values and actual label values are often cross-tabulated into a typical confusion matrix (CM) contained in table 2. Type I 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 II 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.

Our acquaintance with null hypothesis significance testing (NHST) easily allude us to

3. The Proposed Alpheta Ratio

4. Data, Models and Empirical Analysis

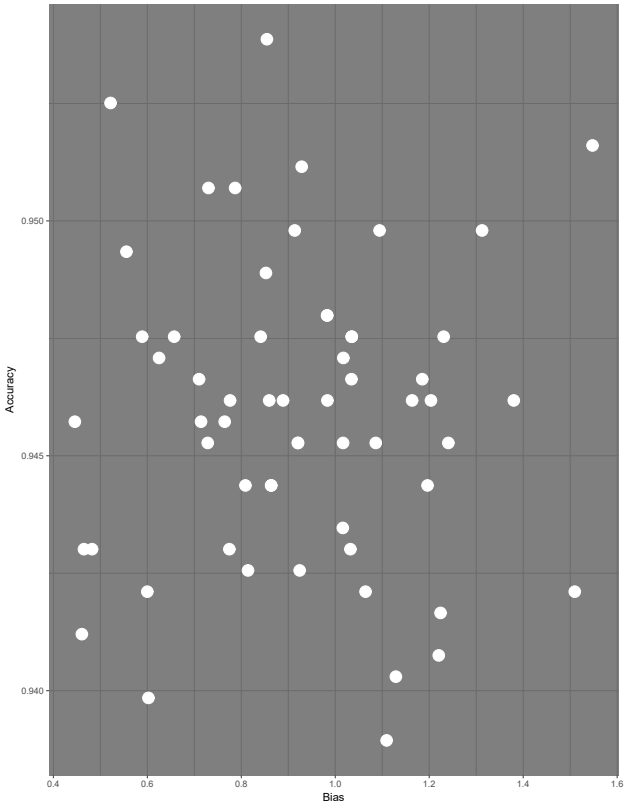
The Phishing Websites Data Set (Mohammad 2015) , used in this work, is available online at the UCI Repository (Lichman 2013). The dataset has been used in several works from the literature

Table 3 Performance Results

Model	Features	Precision	Litmus Ratio
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

Figure 1 Figure Caption.



Note. Text of Notes

that propose techniques for dimensionality reduction and classification. It contains data from 452 patients organized in 16 classes ranging from normal, different types of arrhythmia, and unidentified (see distribution in Table 1). Each instance contains 279 attributes including patients personal data (age, sex, height, and weight), and measurements taken from ECG signals from that patient. Thus, it is clear that the preprocessing of the signals in order to take the measurements was not done in this work

5. Implications

Machine learning hinges on the quality of training data sets that feed into the underlying algorithm. In ML, the more objective the data and the larger the data set, the less possibility of distortion (Rosso 2015).

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.

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).

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.

Lichman M (2013) UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

Mitchell TM (1997) Machine learning.

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

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.