Enhancing Vulnerability Prioritization: Data-Driven Exploit Predictions with Community-Driven Insights

DECISION: accept

AUTHORS: Sasha Romanosky, Jay Jacobs, Octavian Suciu, Benjamin Edwards and Armin Sarabi

Summary of Reviews

- Review 1: 2 (5)
- Review 2: -2 (4)
- Review 3: 2 (3)

Reviews

Review 1

TOTAL SCORE: 2 (accept)
Overall evaluation: 2 (accept)
Reviewer's confidence: 5 (expert)

The paper describes the latest iteration of EPSS, this time trained with more data and latest ML techniques.
Even though the latest model is using ML with many different attributes, the authors do try to provide an explanation for the way that the model works, which attributes are most important.

This is a well-written paper, pleasure to read. Would be even better if the data was made available for others for verification.

Review 2

TOTAL SCORE: -2 (reject)
Overall evaluation: -2 (reject)
Reviewer's confidence: 4 (high)

Overall comments
The paper reports on an effort to collect and analyze software vulnerability exploit data, and build a machine learning model to predict the likelihood of exploitation within 30 days. The model performed significantly better than previous iterations of EPSS, resulting in improved coverage, efficiency, and effort in vulnerability remediation. While the paper is well-structured and easy to read (despite the non-negligible number of typos), I have some concerns about the lack of details in certain section and the impact of certain design choices.

In Section 3.1 of the paper, the collection of exploitation activity is referred to as "ground truth," but it is not actual ground truth. The term "ground truth" typically refers to the actual and definitive facts about a particular phenomenon or situation. However, in the case of the collection of exploitation activity, it is not possible to determine with certainty whether the data sources observed *all* exploits. In particular, the lack of observed exploits for certain vulnerabilities may not necessarily mean that the exploit does not exist. It could be due to the data sources' biases or limitations, which can result in incomplete or inaccurate information. The paper should not refer to the collection of exploitation activity as "ground truth" and instead recognize that the data collected may not be entirely accurate or representative of the actual situation. The paper should assess how this limitation affects the model's accuracy and performance and suggest ways to mitigate any potential biases or inaccuracies in the data.

I also have concerns with Section 3.2, specifically regarding the mapping between the explanation of each block of variables and the actual variables themselves. It appears that the paper is conflating the factors/levels of categorical variables with the actual variables. As a result, for example, it is unclear what the three variables related to social media are after reading this section. To clarify, the paper needs to provide a clear and concise explanation of each variable included in the analysis. The paper should also define any relevant terms or concepts related to the variables to ensure that readers can understand how the variables were measured and how they relate to the research question.
In particulate, the paper should provide a detailed explanation of how the variables were coded and what each code represents. The paper should also discuss any potential limitations or challenges associated with measuring these variables and how they may affect the validity and reliability of the analysis.

The selection of the model is not scientifically grounded. The decision to focus solely on a single ML algorithm, in this case, XGBoost, raises some concerns about the validity and reliability of the results. While it is true that testing all models is impractical, it is essential to consider other models and their potential performance in comparison to XGBoost. Limiting the analysis to a single ML algorithm can lead to the risk of algorithm bias, which means that the model's predictions may be skewed due to the algorithm's limitations and biases. Additionally, this approach may overlook the strengths and weaknesses of other models that may have been better suited for the dataset. While gradient boosted trees have shown good performance on tabular data, this may not be the case for all datasets. ML algorithms' performance often varies depending on the dataset's characteristics, such as the number of features and the amount of data available.

Similarly, the training and evaluation of the model is not fully detailed. For instance, while 5-fold cross-validation may be a reasonable choice, the paper does not consider the limitations and potential bias associated with this choice. It is also essential to perform sensitivity analyses to evaluate the impact of different numbers of folds on the results and ensure that the choice of the number of folds is appropriate for the dataset and research question at hand. Why not choosing 10? Given the size of the dataset, I am not sure the computation argument to choose a low number of folds makes sense.

In short, I think the paper has quite some potential but the different design decisions keep leading to suboptimal scoring systems. I am afraid that we might be having a new version of EPSS each time this SIG trains a different ML algorithm that outperforms the previous one.

Specific comments
- Formatting: it is unclear the meaning of the paragraphs within a framed box.
- Grammar: “a scoring systems” should be “a scoring system.”
- Grammar: “a low barrier to entry for adoption” should be “a low barrier for adoption.”
- Intro: I would avoid using the term “time of this writing” as that is unknown to the reader.
- Intro: The characteristics of the members of this SIG are too broad. It would be interesting to quantify a bit more the demographics of the members. The lack of this information raises questions such as: are all continents equally represented? is some type of industry overrepresented/underrepresented? Given that the paper is setting as requirement that the new scoring system “must address the requirements of practitioners” it is crucial to know which type of practitioners participate in the SIG.
- Intro: Unclear why the paper refers to *state-of-the-art* ML and what would be the comparison against traditional ML. XGBoost is from 2014.
- Section 2: there is no clear information about the key limitations of the previous SIG that hinder adoption. This is essential to understand why a new scoring system was needed.
- Section 2: when referring to performance improvement of the classifier, it is unclear what metric this improvement refers to.

- Section 3.1: “Shadow Server” should be “Shadowserver.”
- Section 3.2: the “social media” variables seem to be a proxy for popularity. I would suggest renaming these.
- Section 4.1: what is the “size of the exposure?”
- Typo: “over-fitting, We” should be “over-fitting, we.”

Review 3

TOTAL SCORE: 2 (accept)
Overall evaluation: 2 (accept)
Reviewer’s confidence: 3 (medium)

The paper presents an improved version of the EPSS scoring system and presents its performance in predicting exploitation in a 30 days window. The paper then compares the efficiency and coverage of different remediation strategies based on the improved versions vs the previous version and CVSS.

The paper tries to solve the challenging problem of addressing vulnerabilities that matter by balancing effort, coverage, and efficiency
The analysis of different remediation strategies clearly shows the problem of ineffective metrics for addressing vulnerabilities.

As a minor comment, given that ground truth data comes from network-based data, is modeling some parts of the CVSS base metric (e.g. attack vector) necessary? This is not really a problem, in particular, if new data from other sources can cover that, but maybe influence the variable importance values in Fig.7 (e.g. for AV).