International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2283125

Luster Regained : A Novel Cyber Incident Risk Prediction Model Using Machine Learning

Meghna "Chili" Pramoda*, Siona "Dolly" Pramoda, Zacha M. Ortiz Correa
Baldwin School of Puerto Rico, Bayamon, PR, USA

Article Info

Publication Issue :
Volume 8, Issue 4
July-August-2022

Page Number : 01-19

Article History :
Accepted: 20 June 2022
Published: 04 July 2022

ABSTRACT

Physical isolation during the COVID-19 pandemic prompted a 45% increase in digital use [5,15], leading to an increase in cyber incidents. This project seeks to understand the risk impact of prolonged internet use and evaluate opportunities for cyber education to lower such risk. In preparation for subsequent work, the project will learn about patterns in distress and the recovery of affected individuals. A 20-question English-language survey (Appendix A) was completed by 6th through 12th graders (n=1,869) across 4 countries.

Analysis of the survey [1, 8, 10, 11, 13] indicated that the number of hours of internet use was a driver of the risk of cyber incidents. In addition to statistical analysis, the methodology used Google's VertexAI AutoML [6] to generate an ensemble model to predict risk (on n=1 basis) from usage patterns (length of usage, gaming use, etc.). The cyber risk predictor model set has high overall accuracy (f1 score of 0.88) and precision and recall of 0.878. This low-cost approach to personalized risk scores could support periodic evaluation and trending of educational effectiveness in cyber safety. Separately, participants reported a strong association (Spearman's Rho = 0.957) between distress from cyber incidents and recovery time. Among the respondents with high distress experiences, there is an urgent need to design support programs to help them cope.

Keywords : Cybersecurity, Cybersafety Education, Ensemble Risk Model, Digital Safety Curriculum, Cyberbullying, Digital Risk Score Prediction

I. INTRODUCTION

Digital use has been growing exponentially across all age groups around the world in the last decade. This path of digital abundance is fraught with numerous risks for younger participants. Approximately 37% of 12-17 year olds in a ~5,000 student study experienced cyberbullying and 15% admitted to inflicting offense on others [9]. According to the Centers for Disease Control, students who experience cyberbullying are more likely to have trouble adjusting at school and are more likely to have mental health and behavioral problems [2].

Recognizing the severity of digital risks to teenagers, DQInstitute released their Digital Citizenship Inventory [3] developed based on 145k respondents across 30 countries. Beyond digital citizenship, individualized risk scores have the highest potential in engaging at-risk youth. Individualized risk scoring is currently lacking per our literature review.

This study explores drivers, patterns, and trends among middle and high school students who have experienced cyber incidents.

Hypotheses

1. According to the Italian Pediatric Society [12], "the duration of time spent using media devices is a main risk factor [for a cyber incident];" this led to the hypothesis that as the number of hours of internet usage outside of school increases, the risk of a cyber incident also increases.
2. An anti-cyberbullying education program, Cyber Friendly Schools [4], found that "the program was associated with significantly greater declines in the odds of involvement in cyber-victimization and perpetration." Therefore, we wanted to know if this was extendable to other aspects of cyber safety. We hypothesized that education in specialty areas of cyber safety can correspond to reducing the risk of a cyber incident.
3. A 2021 study [14] concluded that "some methods of abuse appeared to affect victims more than others." Relating this to cyber safety, we hypothesized that: the more distress a student feels after a cyber incident, the longer it takes to recover.

II. METHODS AND MATERIAL

1. A literature search led to a decision to collect self-reported information on perception, attitudes, usage, and experience patterns directly from students across multiple countries.
2. Given the paucity of literature around proactive cyber safety measurement tools for the specific target population, three research questions were framed:
   a. Did increased internet use outside of school (number of hours) increase cyber incident risk?
   b. Can education in specialty areas of cyber safety correspond to a reduction in the risk of a cyber incident? More broadly, can we predict the factors that have the greatest effect size on the likelihood of the occurrence of a cyber incident?
   c. Do students who feel more distressed from a cyber incident take longer to recover?
3. Data Collection:
   a. An anonymous 20-question English language survey (using SurveyMonkey) was calibrated using readability measurement tools from Readable.com [16] for a Gunning Fog index of 7.5 (ideal for readability in the target age group), with a balanced tone skewing slightly formal, achieving neutrality on sentiment, and slight personalism in language. Refer Appendix A for survey questions.
   b. Respondents were required to respond to 15 out of the 20 questions. The remainder of the questions pertained to sensitive topics such as cyber incident experience and were left optional according to the advice of educators and child psychologists.
   c. The questions addressed broad categories such as
      - ✓ Technology use patterns
      - ✓ Internet use patterns
      - ✓ Hours of internet use
      - ✓ Attitudes towards cyber safety
      - ✓ Perceptions about their skill level to handle cyber safety issues
      - ✓ Cyber incident experiences and help seeking behavior
      - ✓ Prior education in cyber safety related topic areas

- ✓ Education preferences regarding cyber safety related topic areas

4. Survey Responses

a. A total of 9 schools were chosen across 4 countries and appropriate approval was obtained. Using SurveyMonkey, the questionnaire was administered to 6th to 12th graders at 9 schools across 4 countries between Dec 1, 2021 and Feb 10, 2022.

b. **Approvals:** Approval was obtained from school administrators and educators. Under an IRB approval, the survey was distributed through them or their designees. Since the survey was anonymous, the researchers had no contact with the respondents and the survey instrument does not persist individually identifiable information or sensitive data. The questionnaire was administered to 6th to 12th graders between Dec 1, 2021 and Feb 10, 2022.

5. Data Analysis

a. Only completed surveys were admitted into the data analysis.

b. **Analytical Tools**

- ✓ JASP (version 0.16.0.0) [7] for descriptive statistics, regression analysis, pairwise correlations.
- ✓ Microsoft Office Suite & Power BI: for data compilation, data encoding (machine learning preparation), and visualization purposes. Data encoding process is explained in Appendix B1.
- ✓ Google's VertexAI for feature selection, for machine learning model building, validation and testing.

c. Variable Selection

- ✓ Dependent Variable: Boolean of student cyber safety compromise experience (language, inappropriate content, violence, financial theft, etc.), a.k.a "risk of cyber incident" was set as target variable ("dependent variable").
- ✓ Control Variables: All survey participants were (i) enrolled full time in middle or high schools, (ii) primarily in urban areas, (iii) where English is the primary language of instruction, (iv) with school sponsored internet connectivity.
- ✓ Independent Variables: Based on survey questions asked, several of the fields were chosen as independent variables of interest. For further details, see Appendix B2.

d. Analytical Methods

- ✓ Descriptive statistics were used to understand the demographics, technology use, internet use, attitudes and behaviors.
- ✓ Descriptive statistics were used to evaluate participants with cyber incidents and compare them to participants who did not report cyber incidents.
- ✓ Using JASP for regression analysis, tested whether (a) As the number of hours of internet use increases, risk of cyber incident increases (used Wald statistic and p-value for confirmation), and (b) The more distress a student feels after a cyber incident, the longer it takes to recover (statistical significance test).
- ✓ Using JASP, examined the pairwise relationship between each independent variable and dependent variable for statistical significance.
- ✓ Using Google's Vertex AI, encoded survey responses were passed through the AutoML ensemble model building tool to identify the relative importance that meaningfully increase a students' risk for a cyber incident. Output of vertex AI was a predictive classification model. Model efficacy was evaluated using the confusion matrix.

III. RESULTS AND DISCUSSION

The study seeks to identify the drivers of adverse events in cyberspace ("cyber incident") in this population and understand patterns in risk exposure. Whereas digital participation has come to be a permanent part of the lives of youth, results from this study should guide prioritization of education,

advocacy, and policy making to improve cyber safety among minors. As a secondary outcome, the study may offer insights into preventative measures that could be undertaken by families and educational institutions. Further, understanding student patterns in help seeking behaviors could help in the design and activation of safe spaces where impacted students might connect with parents, educators and mental health professionals to develop individual coping skills.

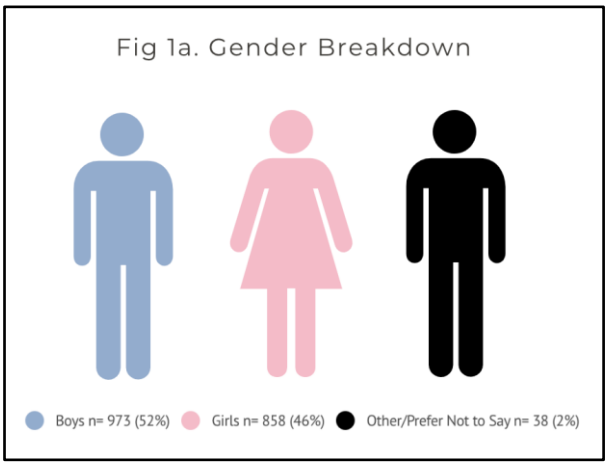
Table 1 : Survey Summary

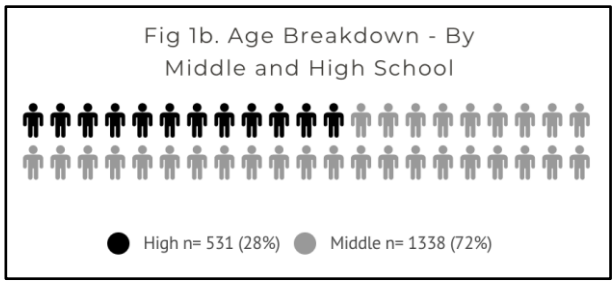
Characteristic	Value
Total Surveyed	2,280
Completed Surveys	1,869 (82%)
Average Response Time	6 minutes 53 seconds
Cyber Incident Rate	20.8%
Average Reported Use of the Internet Outside School	4.5 hours

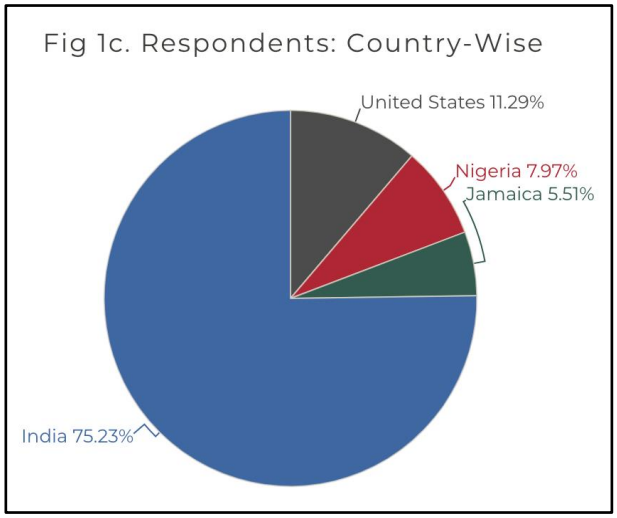
Survey Deep Dives

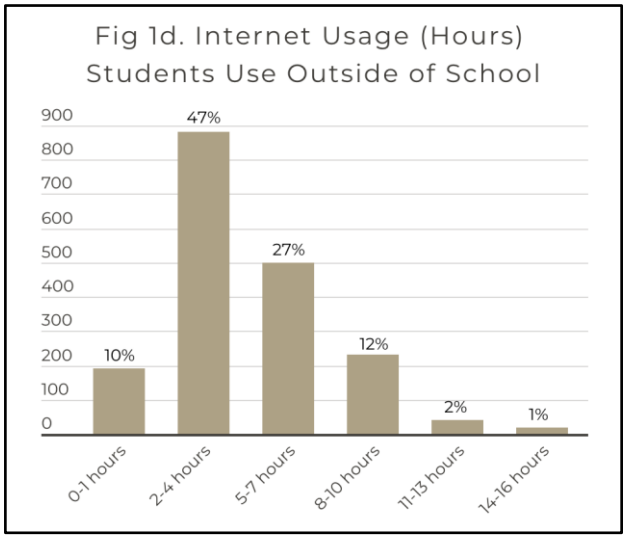
- ✓ 43% of male students and 42% of female students reported using the Internet for 5 or more hours after school. Among the age bins (10-11, 12-13, 14-15, 16-17) of 10-17 year olds, reported high use of the internet (5 hours or more outside school) ranges between 39% and 44%.
- ✓ 60% of male respondents vs. 47% of female respondents reported that they felt highly safe (7 or higher on a scale of 1-10) online.
- ✓ 70% of male respondents vs. 61% of female respondents reported high (7 or higher on a scale of 1-10) confidence in their ability to handle online risk.

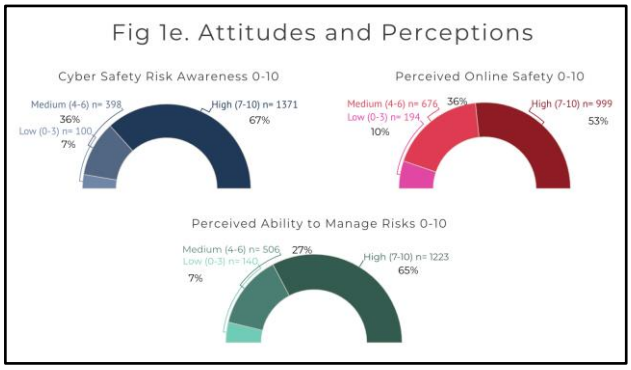
The data collected from the survey is summarized in the figures below (Fig 1a-1e).











Hypothesis 1 - Did increased internet use outside school (# of hours) increase cyber incident risk?

Yes, survey results confirm a statistically significant relationship between increased internet use outside school and a corresponding increase in cyber incident risk (p-value < 0.001, logistic regression).

The respondents were asked to self-report their internet usage outside school hours - on average, students spent about 4.5 hours per day online outside school. To test the hypothesis, a logistic regression model was run, with the risk of cyber incident (binary outcome) as the dependent variable, and the hours of internet use outside school as the independent variable.

Discussion: The logistic regression (fig 2a) noted a strong positive relationship (high Wald score) between the two and that the effect was statistically significant (p value < 0.001, fig 2b). It is understood the small, yet positive odds ratio for the intercept to mean that there is an inherent low risk in the system. And, as hours of internet usage outside school increase, the odds (probability of an event happening / probability of the event not happening) of a cyber incident increase. While the Italian Pediatric Society [12] found, "the duration of time spent using media devices is a main risk factor [for a cyber incident]," this study confirmed this finding across multiple nationalities and geographies (4 countries, 9 schools).

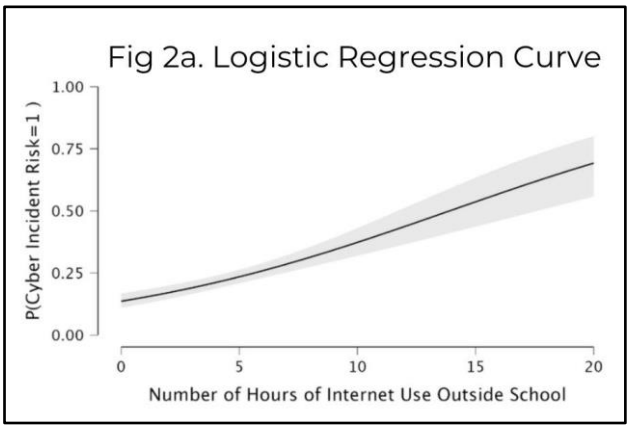


Fig 2b. Logistic Regression Curve Coefficients

Coefficients					
				Wald Test	
	Estimate	Standard Error	Odds Ratio	Wald Statistic	p
Intercept	-1.99	0.112	0.137	315.376	< 0.001
hours_internet_use_outside_school	0.137	0.019	1.147	52.459	< 0.001

Hypothesis 2 : Can education in specialty areas of cyber safety correspond to a reduction in the risk of a cyber incident?

Yes, there is a statistically significant (with low correlation, p-value < 0.05) relationship between education in specialty areas of cyber safety and a corresponding reduction in cyber incidents.

A. A complex problem is not completely explained by one feature.

The figure below (fig 3a) shows features that are positively and negatively correlated with risk of a

cyber incident (p-value < 0.05). However, the pairwise correlations are low (-0.141 to 0.166). This suggests that one of these features is not an overwhelming driver of the outcome. As such, a combination of these features may better explain the dependent variable.

Fig 3a. Pairwise Correlation With risk_experience

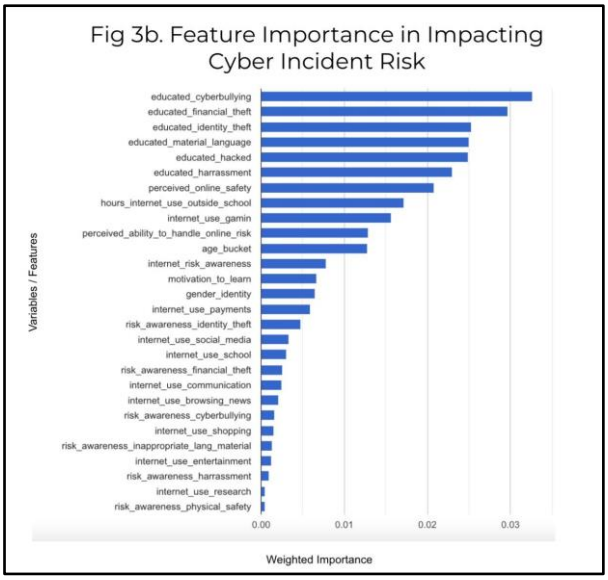
Feature	Estimate	Odds Ratio	p	Correlation
hours_internet_use_outside_school	0.115	1.122	< .001	0.166
risk_awareness_other	0.948	2.58	< .001	0.102
educated_material_language	0.615	1.849	< .001	0.022
tech_use_public_wifi	0.597	1.816	0.001	0.152
risk_awareness_inappropriate_lang_material	0.584	1.794	0.007	0.052
internet_use_social_media	0.387	1.473	0.02	0.117
internet_use_payments	0.371	1.449	0.026	0.07
perceived_online_safety	-0.133	0.875	< .001	-0.096
education_pref_classroom_instruction	-0.573	0.564	< .001	-0.141
internet_use_school	-0.735	0.479	0.002	-0.075
age_bucket	-0.183	0.833	0.01	-0.033
internet_risk_awareness	-0.085	0.919	0.013	-0.1
educated_hacked	-0.401	0.669	0.028	-0.109
education_pref_peer_discussion	-0.27	0.763	0.048	-0.059

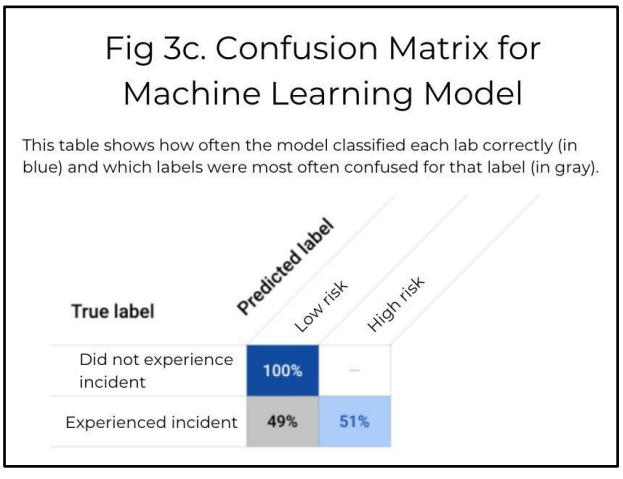
The p-value of all the rows in fig 3a is < 0.05 indicating that each correlation, while small, is statistically significant.

Not all specialty areas of cyber safety education were found to be statistically significant. To understand the impact and relative importance of education in specific domains of cyber safety, a machine learning approach was used, and feature importance was identified. The target variable continued to be set as the risk of cyber incident occurring. All features that constituted information available prior to the occurrence of a cyber incident were included for ranking. Fig 3b shows the relative importance of each feature in predicting the risk of a cyber incident. Whereas pairwise correlation (fig 3a) shows direction, feature importance (fig 3b) quantifies relative effect size.

Feature importance looks at the relative importance of each of the features as it pertains to the impact on the dependent variable (risk of cyber incident). Six out of the top ten features (i) education on cyber bullying, (ii) education on financial theft, (iii) education on identity theft, (iv) education on appropriate language, (v) education on hacking, and (vi) education on harassment, are education related. This indicates that instituting an education program that addresses these knowledge areas may meaningfully reduce risk of cyber incidents.

Two among the top ten features, (i) perceived online safety and, (ii) perceived ability to handle online risk are elements of personal experience that may be partially addressed through investments in resilience training. Across eight out of top ten features, it appears possible to offset some of the risks associated with increased hours spent online outside school and internet use for gaming.





B. Predicting Risk of Cyber Incidents

About 1 in 5 survey respondents revealed that they'd experienced a cyber incident (ground truth / training set / survey results). The goal was to build a model to help identify students at high risk of a cyber incident. Predicting such an incident before it occurs would pave the way to (a) minimize its chances through education, and, (b) provide culturally appropriate support in its aftermath. Even though the pairwise correlations between the features and target variable (risk_experience) were low, corresponding low p-values indicated that the data was stable. An approach to take into account multiple features at once and predict at the n=1 level was necessary to follow up with those interventions.

The data was encoded, then split into training, validation, and test sets (80/10/10) in Google's machine learning suite, Vertex AI, prior to starting the AutoML model training process. The goal was to predict the risk of a cyber incident before it occurred. Feature importance (fig 3b), precision, recall, f1 score, and model efficacy (confusion matrix, fig 3c) were reviewed.

The confusion matrix sheds light upon the efficacy of the model in accurately predicting true positives and false negatives. The model's precision and recall were 0.878, overall accuracy (f1 score) was over 88%. Digging a little deeper, the model was correct every single time it predicted that a student was a "1" (meaning the student was at high risk for a cyber incident). When the model predicted a student was not a high risk though, it was right 2 out of 3 times. This means that even though the model needs to be improved to help accurately identify high risk students, we derived some benefit from it - instead of finding a high risk student through counselors interviewing 5 students (20.8% incidence reported in this survey), the counselors should expect to find 1 high risk student for every three they interview. This allows for more directed efforts in supporting students with highest need.

C. Model Improvement

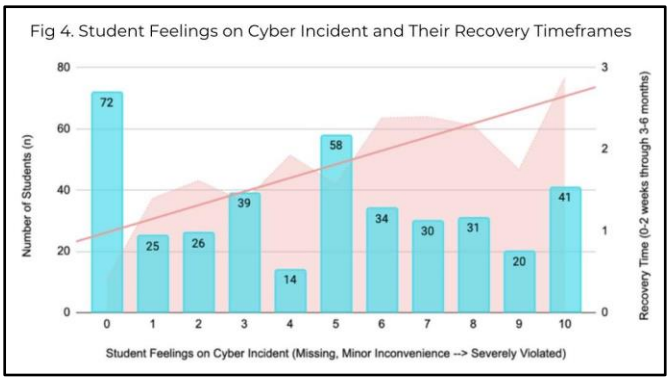
- **Increase Dataset Size to Address Class Imbalance:** The training dataset is unbalanced (cyber incidents are reported in ⅕ students). Typically, the model developer would apply techniques to synthetically augment the training data to reduce class imbalance. Since this training set is derived from a survey, this method of addressing class imbalance is not advisable. Instead, the model can be improved once the survey is rolled out to more schools (and a lot more respondents fill out the survey).
- **Noise Reduction in Features:** Identify opportunities to consolidate closely related features (for example, the different facets of education into a single education feature) for model efficiency and noise reduction purposes.
- **Expand Survey:** Enhance survey to pursue new features to improve prediction accuracy.

Hypothesis 3: Do students who feel more distressed from a cyber incident take longer to recover?

Yes, the survey analysis suggests that students who feel more distressed from a cyber incident take longer to recover from it ($p<0.001$, linear regression).

Among the 390 (20.8% of N of 1,869) respondents who reported prior experience with a cyber incident, 18% (72 respondents) of students who experienced cyber incidents abstained from sharing how it made them feel and how long it took them to recover from it (Recovery time: categoric variable with possible responses ranging from 1: 0-4 weeks 2: 1-6 month and 3:6 months or longer). Among the respondents that chose to share such information, there is a statistically significant relationship ($p<0.001$, linear regression) between student feelings on cyber incidents and how long it took them to recover from the experience.

Fig 4 confirms our hypothesis that the more burdened students feel about cyber incidents they experience, the longer it takes them to recover from it. Given that >30% of respondents indicate higher levels of distress and >1 month of recovery time, further research is warranted to understand and design specific intervention for such individuals. It is vitally important for such students to receive support and intervention from qualified behavioral health / medical professionals.



Model Improvement

Survey responses on time to recover were categorical in nature, and in order to create a linear regression, numeric responses were assigned to the range of responses (1: 0-4 weeks 2: 1-6 month and 3: 6 months or longer). Equal-sized time buckets, or a slider to allow respondents to choose the actual number of days to recover between 0 and 180 might lead to a better linear regression model.

Limitations

1. Survey was in English language, but not all participants may be native speakers of English. Survey was administered in 4 countries. The responses collected are geographically unbalanced and not representative of the world population. Expansion of the survey response collection across the globe and increasing response volume would help develop a more representative voice of the global teenager.
2. Survey was administered in multiple settings (e.g., classroom, assigned by teacher as homework, requested by student ambassador as part of non-academic club participation) which may have created variance across responses. Students may have engaged in conversation or shared opinions that could have influenced their survey responses.
3. Given that this is a survey based study, causation cannot be established.
4. Certain groups of respondents demonstrate a tendency to skew high in their responses to categorical questions. It is likely that such predisposition may be a function of cultural norms that favor confidence or optimism and stigmatize information sharing on adversity. For instance, the responses indicate a substantially lower prevalence of cyber incidents (20.8%) in spite of defining the issue more broadly compared to 37% in Patchin [9] that was limited to cyberbullying.
5. Imbalanced classes in ground truth to train machine learning models leads to mixed model performance. While in case of image data, it is possible to overcome this limitation through augmenting the training set with synthetic / transformed images, in survey data, this is not possible.

IV. CONCLUSION

Study results demonstrate a clear relationship between externally addressable, individually addressable and non-addressable factors and risk of cyber incidents.

1. There was a strong positive relationship between the number of hours of internet usage outside school and the risk of a cyber incident. However, the pairwise correlation is weak and suggests that cyber incident risk is impacted by other features beyond prolonged internet usage outside school.

2. Education on inappropriate language and materials is associated with increased risk of cyber incidents. Further study is necessary to qualify whether there is a causal relationship or a spurious correlation.
3. As individual features, (i) classroom instruction, (ii) education on hacking, and (iii) peer discussion are associated with reducing risk of cyber incidents. The cyber risk predictor model set has high overall accuracy, precision and recall. This low-cost approach to personalized risk scores could support periodic evaluation and trending of educational effectiveness in cyber safety.
4. Multiple education related features can be deployed to impact risk of cyber incidents. In continued exploration of this topic, quantifying the actual effects of comprehensive educational efforts could support policy making and guide program rollouts in schools and communities.
5. High correlation between students' experience of distress and recovery time as well as a high reported prevalence of deep distress points to an urgency to understand, analyze and design support for such individuals. Studying this in the future could develop models for early identification and matching of such students with behavioral health / clinical experts.
6. This is a global study. To develop a globally relevant model, data collection efforts need to be expanded to collect a wider dataset with culturally appropriate questions that will pave the road for data-informed personalization of educational content for youth worldwide.

V. ACKNOWLEDGEMENTS

This research is part of the advocacy and education efforts at SafeTeensOnline (a youth-led global cyber safety initiative). This study is possible with guidance and approval from (and special thanks to) The Baldwin School of Puerto Rico's IRB committee (Ms. C. Rivera, Ms. V. Banks, Ms. M. del Llano, and Ms. L. Muniz), mentor (Ms. K. Srinivas) and expert reviewers (Dr. D. Varadarajan and Dr. V. Sundaram).

VI. REFERENCES

[1]. 11 facts about cyberbullying. (n.d.). DoSomething.Org. https://www.dosomething.org/us/facts/11-facts-about-cyber-bullying (Accessed February 12, 2022).

[2]. CDC. (2019). #StopBullying. Centers for Disease Control and Prevention. https://www.cdc.gov/injury/features/stop-bullying/index.html (Accessed December 29, 2021).

[3]. Collaborative R&D – DQ institute. (n.d.). Retrieved February 23, 2022, from https://www.dqinstitute.org/collaborative-rd/#st

[4]. Cross et al. (2015). Longitudinal impact of the Cyber Friendly Schools program on adolescents' cyberbullying behavior. Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21609 (Accessed October 3, 2021).

[5]. Enough is enough: Internet safety. (n.d.). Enough.Org. Retrieved February 12, 2022, from https://enough.org/stats_internet_safety

[6]. Google's Vertex AI software for machine learning model building and evaluation. https://cloud.google.com/vertex-ai (Accessed February 20, 2022).

[7]. JASP open-source software for statistical analysis. https://jasp-stats.org/ v0.16.1(Accessed February 20, 2022).

[8]. Kemp, S. (2021, January 27). Digital 2021: Global overview report — datareportal – global digital insights. DataReportal – Global Digital

Insights. https://datareportal.com/reports/digital-2021-global-overview-report (Accessed September 29, 2021).

[9]. Patchin, B. J. W. (2019, July 9). 2019 cyberbullying data. Cyberbullying Research Center. https://cyberbullying.org/2019-cyberbullying-data (Accessed September 17, 2021).

[10]. Patchin, J. W., & Hinduja, S. (2020, December 18). Tween statistics (9- to 12-year-olds). Cyberbullying Research Center. https://cyberbullying.org/tween-statistics (Accessed September 14, 2021).

[11]. Przybylski, A. K., & Nash, V. (2018). Internet filtering and adolescent exposure to online sexual material. Cyberpsychology, Behavior and Social Networking, 21(7), 405–410. https://doi.org/10.1089/cyber.2017.0466 (Accessed October 11, 2021).

[12]. Spina, G., Bozzola, E., Ferrara, P., Zamperini, N., Marino, F., Caruso, C., Antilici, L., & Villani, A. (2021). Children and adolescent's perception of media device use consequences. International Journal of Environmental Research and Public Health, 18(6). https://doi.org/10.3390/ijerph18063048 (Accessed October 3, 2021).

[13]. Thorn & Benenson Strategy Group. (2021, May). Responding to Online Threats: Minors' Perspectives on Disclosing, Reporting, and Blocking - Findings from 2020 quantitative research among 9–17 year olds. Thorn.Org. https://info.thorn.org/hubfs/Research/Responding%20to%20Online%20Threats_2021-Full-Report.pdf (Accessed October 26, 2021).

[14]. Vakhitova, Z. I., Alston-Knox, C. L., Reeves, E., & Mawby, R. I. (2021). Explaining victim impact from cyber abuse: An exploratory mixed methods analysis. Deviant Behavior, 1–20. https://doi.org/10.1080/01639625.2021.1921558 (Accessed October 3, 2021).

[15]. Watson, A. (2020, June 18). Media usage during COVID-19 by country. Statista. https://www.statista.com/statistics/1106498/home-media-consumption-coronavirus-worldwide-by-country/ (Accessed February 21, 2022).

[16]. Take control of your content with ReadablePro. (2018, January 18). Readable. https://readable.com/ (Accessed November 23, 2021).

I. APPENDICES

Appendix A - Survey

We are middle and high school students creating awareness about safety and security while communicating online. We are collecting preliminary information in order to understand the need and basic knowledge in this area using this survey. This anonymous survey is applicable to students who are either in middle school, high school, or college. The survey results will be used only to assist us with the awareness campaign. Thank you for your time.

* 1. Which one best describes you?

Asian or Pacific Islander | Black or African American | Hispanic or Latino | Native American or Alaskan Native | White or Caucasian | Multi-racial or bi-racial | A race or ethnicity not listed here

* 2. Which year were you born in?

2002 - 2003 | 2004 - 2005 | 2006 - 2007 | 2008 - 2009 | 2010 - 2011

* 3. Which best describes your gender identity?

Male | Female | Other | Prefer not to say

* 4. What technology do you use? (check all that apply)

Desktop or laptop computer | Home Internet (WiFi) | Smartphone | Tablet | Public WiFi | Gaming consoles (X-box/Play Station/Nintendo Switch etc.)

* 5. What do you use internet / connected device for? (check all that apply)

Social media | School | Communication - texting, email, video | Entertainment - music, streaming, sports | Shopping | Gaming | Payments - banking, cashless transfer
Browsing / news | Research - find information | Other (please specify)

* 6. How many hours outside school do you spend online each day?

0-16

* 7. On a scale of 1 to 10 (1 being the least safe, 10 being the most safe) how would you rate your awareness around online and digital security risks?

0 - 10

* 8. On a scale of 1 to 10 (1 being the least safe, 10 being the most safe), how safe do you feel in the online and digital world?
0 - 10

* 9. On a scale of 1 to 10 (1 being "I feel helpless" to 10 being "I am in control"), how well do you think you can handle online and digital risk?
0 - 10

* 10. Are you aware of any of the following risks?

Exposure to inappropriate material/language	Yes	No
Cyber bullying / reputational harm	Yes	No
Financial theft (people stealing your credit cards, or money)	Yes	No
Identity theft (people impersonating you and causing harm)	Yes	No
Harassment - online predator, stalking, sexual harassment, threats	Yes	No
Risk of physical safety	Yes	No
Other (please specify)		

* 11. Have you experienced any of the above?

Yes or No

12. If yes, which of these did you experience? (Check all that apply)

Exposure to inappropriate material/language
Cyber bullying / reputational harm
Financial theft (people stealing your credit cards, or money)
Identity theft (people impersonating you and causing harm)
Harassment - online predator, stalking, sexual harassment, threats
Risk of physical safety
Other (please specify)

13. On a scale of 1 to 10 (1 being “no big deal” to 10 being “this is the worst thing ever”), how did it make you feel? (accept whole numbers 1-10)

0 - 10

14. Who did you seek help from? (Check all that apply)
Parent | Friend | Teacher | Counselor | I did not seek help | Other (please specify)

15. How long did it take you to recover from the negative impact of the adverse event?
0-2 weeks | 2-4 weeks | 1-3 months | 3-6 months | > 6 months | Not yet recovered

16. What changes have you implemented since the adverse event(s)? (check all that apply)
Learned more about digital risks | Learned about how to implement digital security measures | Implemented some digital security measures | Implemented comprehensive digital security measures | Reduced

my digital usage/footprint | Have not changed anything

* 17. Have you been educated on any of these topics through your school?

Exposure to inappropriate material/language	Yes	No
Cyber bullying / reputational harm	Yes	No
Financial theft (people stealing your credit cards, or money)	Yes	No
Identity theft (people impersonating you and causing harm)	Yes	No
Harassment - online predator, stalking, sexual harassment, threats	Yes	No
Risk of physical safety	Yes	No
Other (please specify)		

* 18. How would you like to be educated on digital safety and methods? (Check all that apply)

Reading materials | Watching videos | Classroom instruction (teacher) | Presentation and discussion hosted by peers / friends | Other (please specify)

* 19. On a scale of 1 to 10 (1 being the least motivated and 10 being the most motivated), how motivated are you to learn more about risks and methods in digital safety?

0 - 10

* 20. What is stopping you from being safe online? (check all that apply)

I don't have time
It is not cool | I don't know how to do it | I don't believe it is possible to be safe | I don't believe it is important | I don't have access to tools | Other (please specify)

Appendix B1 - Data Encoding

Variable Name	Question	Encodings
race	Which one best describes you?	0 = White or Caucasian 1 = Black or African American 2 = Hispanic or Latino 3 = Asian or Pacific Islander 4 = Native American or Alaskan Native 5 = Multi-racial or bi-racial 6 = A race or ethnicity not listed here
age_bucket	Which year were you born in?	4 = 2002 - 2003 3 = 2004 - 2005 2 = 2006 - 2007 1 = 2008 - 2009 0 = 2010 - 2011
gender_identity	Which best describes your gender identity?	1 = Male 2 = Female 3 = Other 4 = Prefer not to say

tech_use_desktop	What technology do you use? (check all that apply) | Desktop or laptop computer	1 = TRUE 2 = FALSE
tech_use_home_wifi	What technology do you use? (check all that apply) | Home Internet (WiFi)	1 = TRUE 2 = FALSE
tech_use_smartphone	What technology do you use? (check all that apply) | Smartphone	1 = TRUE 2 = FALSE
tech_use_tablet	What technology do you use? (check all that apply) | Tablet	1 = TRUE 2 = FALSE
tech_use_public_wifi	What technology do you use? (check all that apply) | Public WiFi	1 = TRUE 2 = FALSE
Variable Name	**Question**	**Encodings**
tech_use_gaming_consoles	What technology do you use? (check all that apply) | Gaming consoles (X-box/Play Station/Nintendo Switch etc.)	1 = TRUE 2 = FALSE
tech_use_other	What technology do you use? (check all that apply) | Other (please specify)	1 = TRUE 2 = FALSE
internet_use_social_media	What do you use internet / connected device for? (check all that apply) | Social media	1=Yes 2 = No
internet_use_school	What do you use internet / connected device for? (check all that apply) | School	1=Yes 2 = No
internet_use_communication	What do you use internet / connected device for? (check all that apply) | Communication - texting, email, video	1=Yes 2 = No
internet_use_entertainment	What do you use internet / connected device for? (check all that apply) | Entertainment - music, streaming, sports	1=Yes 2 = No
internet_use_shopping	What do you use internet / connected device for? (check all that apply) | Shopping	1=Yes 2 = No
internet_use_gamin	What do you use internet / connected device for? (check all that apply) | Gaming	1=Yes 2 = No
internet_use_payments	What do you use internet / connected device for? (check all that apply) | Payments - banking, cashless transfer	--
internet_use_browsing_news	What do you use internet / connected device for? (check all that apply) | Browsing / news	1=Yes 2 = No
internet_use_research	What do you use internet / connected device for? (check all that apply) | Research - find information	1=Yes 2 = No
internet_use_other	What do you use internet / connected device for? (check all that apply) | Other	1=Yes 2 = No
internet_use_other_freetext	What do you use internet / connected device for? (check all that apply) | Other (please specify)	99 = free text

hours_internet_use_outside_school	How many hours outside school do you spend online each day?	

Variable Name	Question	Encodings
internet_risk_awareness	On a scale of 1 to 10 (1 being the least safe, 10 being the most safe) how would you rate your awareness around online and digital security risks?	
perceived_online_safety	On a scale of 1 to 10 (1 being the least safe, 10 being the most safe), how safe do you feel in the online and digital world?	
perceived_ability_to_handle_online_risk	On a scale of 1 to 10 (1 being "I feel helpless" to 10 being "I am in control"), how well do you think you can handle online and digital risk?	
risk_awareness_inappropriate_lang_material	Are you aware of any of the following risks? | Exposure to inappropriate material/language	1=Yes 2 = No
risk_awareness_cyberbullying	Are you aware of any of the following risks? | Cyber bullying / reputational harm	1=Yes 2 = No
risk_awareness_financial_theft	Are you aware of any of the following risks? | Financial theft (people stealing your credit cards, or money)	1=Yes 2 = No
risk_awareness_identity_theft	Are you aware of any of the following risks? | Identity theft (people impersonating you and causing harm)	1=Yes 2 = No
risk_awareness_harrassment	Are you aware of any of the following risks? | Harassment - online predator, stalking, sexual harassment, threats	1=Yes 2 = No
risk_awareness_physical_safety	Are you aware of any of the following risks? | Risk of physical safety	1=Yes 2 = No
risk_awareness_other	Are you aware of any of the following risks? | Other (please specify)	1=Yes 2 = No
risk_experience	Have you experienced any of the above?	1=Yes 2 = No
risk_experience_inappropriate_materia_language	If yes, which of these did you experience? (Check all that apply) | Exposure to inappropriate material/language	1=Yes 2 = No
risk_experience_cyberbullying	If yes, which of these did you experience? (Check all that apply) | Cyber bullying / reputational harm	1=Yes 2 = No

Variable Name	Question	Encodings

risk_experience_financial_theft	If yes, which of these did you experience? (Check all that apply) | Financial theft (people stealing your credit cards, or money)	1=Yes 2 = No
risk_experience_identity_theft	If yes, which of these did you experience? (Check all that apply) | Identity theft (people impersonating you and causing harm)	1=Yes 2 = No
risk_experience_harrassment	If yes, which of these did you experience? (Check all that apply) | Harassment - stalking, sexual harassment, threats	1=Yes 2 = No
risk_experience_physical_safety	If yes, which of these did you experience? (Check all that apply) | Risk of physical safety	1=Yes 2 = No
risk_experience_other	If yes, which of these did you experience? (Check all that apply) | Being hacked - Someone took my data, gained control of my device, demanded a ransom etc.	1=Yes 2 = No
risk_experience_feeling	On a scale of 1 to 10 (1 being “no big deal” to 10 being “this is the worst thing ever”), how did it make you feel? (accept whole numbers 1-10)	
help_seek_parent	Who did you seek help from? (Check all that apply) | Parent	1=Yes 2 = No
help_seek_friend	Who did you seek help from? (Check all that apply) | Friend	1=Yes 2 = No
help_seek_teacher	Who did you seek help from? (Check all that apply) | Teacher	1=Yes 2 = No
help_seek_counselor	Who did you seek help from? (Check all that apply) | Counselor	1=Yes 2 = No
help_seek_other	Who did you seek help from? (Check all that apply) | Some person other than ones named above	1=Yes 2 = No
help_seek_did_not_seek	Who did you seek help from? (Check all that apply) | I did NOT seek help	1=Yes 2 = No
help_seek_other_freetext	Who did you seek help from? (Check all that apply) | Other (please specify)	0 = not fileld 99 = free text
Variable Name	**Question**	**Encodings**
recovery_time	How long did it take you to recover from the negative impact of the adverse event?	0 = Not Answered 1 = 0 - 2 weeks 2 = 2 - 4 weeks 3 = 1 - 3 months 4 = 3 - 6 months 5 = > 6 months 6 = Not yet recovered
changes_learn_digrisk	What changes have you implemented since the adverse event(s)? (check all that apply) | Learned more about digital risks	1=Yes 2 = No

changes_learn_digisec_measures	What changes have you implemented since the adverse event(s)? (check all that apply) | Learned about how to implement digital security measures	1=Yes 2 = No
changes_implement_digisec_measures	What changes have you implemented since the adverse event(s)? (check all that apply) | Implemented some digital security measures	1=Yes 2 = No
changes_implement_comp_digisec_measures	What changes have you implemented since the adverse event(s)? (check all that apply) | Implemented comprehensive digital security measures	1=Yes 2 = No
changes_implement_reduce_footprint	What changes have you implemented since the adverse event(s)? (check all that apply) | Reduced my digital usage/footprint	1=Yes 2 = No
changes_implement_none	What changes have you implemented since the adverse event(s)? (check all that apply) | Have not changed anything	1=Yes 2 = No
educated_material_language	Have you been educated on any of these topics through your school? | Exposure to inappropriate material / language	1=Yes 2 = No
educated_cyberbullying	Have you been educated on any of these topics through your school? | Cyber bullying / reputational harm	1=Yes 2 = No
educated_financial_theft	Have you been educated on any of these topics through your school? | Financial theft (people stealing your credit cards, or money)	1=Yes 2 = No

Variable Name	Question	Encodings
educated_identity_theft	Have you been educated on any of these topics through your school? | Identity theft (people impersonating you and causing harm)	1=Yes 2 = No
educated_harrassment	Have you been educated on any of these topics through your school? | Harassment - stalking, sexual harassment, threats	1=Yes 2 = No
educated_physical_safety	Have you been educated on any of these topics through your school? | Risk of physical safety	1=Yes 2 = No
educated_hacked	Have you been educated on any of these topics through your school? | Being hacked – someone took my data, gained control of my device, demanded a ransom etc.	1=Yes 2 = No
educated_other	Have you been educated on any of these topics through your school? | Other (please specify)	
education_pref_reading_materials	How would you like to be educated on digital safety and methods? (Check all that apply) | Reading materials	1=Yes 2 = No

education_pref_watching_videos	How would you like to be educated on digital safety and methods? (Check all that apply) | Watching videos	1=Yes 2 = No
education_pref_classroom_instruction	How would you like to be educated on digital safety and methods? (Check all that apply) | Classroom instruction (teacher)	1=Yes 2 = No
education_pref_peer_discussion	How would you like to be educated on digital safety and methods? (Check all that apply) | Presentation and discussion hosted by peers / friends	1=Yes 2 = No
education_pref_other_freetext	How would you like to be educated on digital safety and methods? (Check all that apply) | Other (please specify)	
motivation_to_learn	On a scale of 1 to 10 (1 being the least motivated and 10 being the most motivated), how motivated are you to learn more about risks and methods in digital safety?	
stopping_no_time	What is stopping you from being safe online? (check all that apply) | I don't have time	
Variable Name	**Question**	**Encodings**
stopping_not_cool	What is stopping you from being safe online? (check all that apply) | It is not cool	
stopping_no_knowldge	What is stopping you from being safe online? (check all that apply) | I don't know how to do it	
stopping_its_impossible	What is stopping you from being safe online? (check all that apply) | I don't believe it is possible to be safe	
stopping_its_unimportant	What is stopping you from being safe online? (check all that apply) | I don't believe it is important	
stopping_no_tools	What is stopping you from being safe online? (check all that apply) | I don't have access to tools	
stopping_other_freetext	What is stopping you from being safe online? (check all that apply) | Other (please specify)	

Appendix B2 - Independent Variables

Demographics
race
age_bucket
gender_identity

Tech/Internet Use
tech_use_desktop
tech_use_home_wifi
tech_use_smartphone
tech_use_tablet
tech_use_public_wifi
tech_use_gaming_consoles
tech_use_other

internet_use_social_media
internet_use_school
internet_use_communication
internet_use_entertainment
internet_use_shopping
internet_use_gamin
internet_use_payments
internet_use_browsing_news
Tech/Internet Use
internet_use_research
internet_use_other
internet_use_other_freetext
hours_internet_use_outside_school

Awareness
internet_risk_awareness
perceived_online_safety
perceived_ability_to_handle_online_risk
risk_awareness_inappropriate_lang_material
risk_awareness_cyberbullying
risk_awareness_financial_theft
risk_awareness_identity_theft
risk_awareness_harrassment
risk_awareness_physical_safety
risk_awareness_other

Learning
educated_material_language
educated_cyberbullying
educated_financial_theft
educated_identity_theft
educated_harrassment
educated_physical_safety
educated_hacked
educated_other
education_pref_reading_materials
education_pref_watching_videos
education_pref_classroom_instruction
education_pref_peer_discussion
education_pref_other_freetext
motivation_to_learn

Post-cyber-incident Behaviors
risk_experience_feeling
help_seek_parent
help_seek_friend
help_seek_teacher
help_seek_counselor
Post-cyber-incident Behaviors
help_seek_other
help_seek_did_not_seek
help_seek_other_freetext
recovery_time
changes_learn_digrisk
changes_learn_digisec_measures
changes_implement_digisec_measures
changes_implement_comp_digisec_measures
changes_implement_reduce_footprint
changes_implement_none
stopping_no_time
stopping_not_cool

stopping_no_knowldge
stopping_its_impossible
stopping_its_unimportant
stopping_no_tools
stopping_other_freetext

Cite this article as :

Meghna 'Chili' Pramoda, Siona 'Dolly' Pramoda, Zacha M. Ortiz Correa, "Luster Regained : A Novel Cyber Incident Risk Prediction Model Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 4, pp. 01-19, July-August 2022. Available at doi : https://doi.org/10.32628/CSEIT2283125
Journal URL : https://ijsrcseit.com/CSEIT2283125