Predicting Data Breaches with Machine Learning: A Focus on Data Type and Vulnerabilities (November 2023)

Y Sugiyama

George Mason University, Fairfax, VA 22030 USA

Corresponding author: Y Sugiyama (e-mail: ysugiyam@ protonmail.com).

ABSTRACT

Data breaches pose challenges for organizations, necessitating a proactive approach to mitigate risks and grasp the underlying dynamics. This study uses machine learning, data science techniques, and NLP to predict Massachusetts data breaches, aiming to uncover patterns and susceptibility. Organizational susceptibility, data protection strategies, and motivations behind data breaches will be investigated. A Decision Tree classifier was employed on a cleaned, transformed, and normalized dataset with balanced representation. The model achieves 81% accuracy in the training dataset and 80% in the test dataset, with only the two majority classes “Finance” and “Corporation”. This research emphasizes the need for comprehensive data collection, improved NLP, and model robustness. Future work will involve finding a more practical way to categorize these organizations using NLP, exploring more machine learning models, hyperparameter tuning, and collecting additional data, such as geographical locations, to further enhance predictive accuracy. Finally, the study emphasizes the importance of understanding and addressing the multifaceted challenges created by data breaches in diverse organizational contexts.

INDEX TERMS

NLP; Machine learning; Data breaches; Decision tree; Cyber vulnerabilities

I. INTRODUCTION

In the realm of our current efforts in cybersecurity, data breaches have become a common threat, with their frequency and impact on the rise. These breaches happen when unauthorized individuals access private information, causing financial losses, damage to reputations, and legal issues for the people and organizations involved. To address this growing problem, experts within this field are adopting sophisticated technologies like machine learning to forecast and prevent data breaches. This research project aims to predict and categorize organized data breaches by looking at the type of personal information they target, such as credit card data or social security numbers. It also seeks to identify the type of organizations, where these breaches are more likely to occur. The motivation for this project comes from the increasing number and severity of data breaches, highlighting the need for accurate predictive methods that consider location-based risk factors. To further explain this motivation, consider how major retail chains have suffered significant financial losses and damage to their reputation due to data breaches. Government agencies also struggle to protect sensitive data in different regions. This emphasizes the need for the precise method of predicting data breaches prior to their occurrence.

In the context of this research, we have identified several critical research questions that will help us better understand data breaches:

1. Are certain types of organizations more prone to data breaches than others?
2. How effective are strategies like Data Encryption and Credit Monitoring in preventing or reducing the impact of data breaches?
3. Can the predictions from our research reveal patterns and motivations behind organized data breaches, offering insights into their targets?

In pursuit of these questions, our project has three primary objectives:

1. To analyze and evaluate the vulnerability of different types of organizations to data breaches, considering factors that make them more susceptible.
2. To assess the effectiveness of data protection strategies, particularly focusing on Data Encryption and Credit Monitoring.
3. To gain a comprehensive understanding of the patterns and motivations behind organized data breaches, providing valuable insights into their targets and how they are executed.

This research project aims to shed light on the complex world of data breaches, equipping stakeholders with the tools and knowledge needed to prevent and mitigate these cyber threats effectively.

II. DATA

The

III. EDA

A graph of a credit and debit

Description automatically generated

FIGURE 1.  A facet plot of the frequency of targeted data from the data breaches

A graph of a number of bars

Description automatically generated with medium confidenceFIGURE 2.  Distribution of breach occurrence by year

A pie chart with text overlay

Description automatically generated

FIGURE 3.  A pie chart of organizations providing credit monitoring

A graph of data encryption status

Description automatically generatedFIGURE 4.  A scaled histogram of organizations with data encryption

A graph of a number of blue squares

Description automatically generated with medium confidence FIGURE 5.  A scaled histogram of data breaches where a mobile device was lost or stolen.

IV. MODEL

The Model

V. MODEL ANALYSIS AND EVALUATION

The Results

TABLE 1

Training Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Corporation | 0.79 | 0.81 | 0.80 | 3434 |
| Finance | 0.81 | 0.78 | 0.79 | 3434 |
| Accuracy |  |  | 0.80 | 6868 |
| Macro avg | 0.80 | 0.80 | 0.80 | 6868 |
| Weighted avg | 0.80 | 0.80 | 0.80 | 6868 |

TABLE 2

Testing Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Corporation | 0.67 | 0.82 | 0.74 | 1164 |
| Finance | 0.88 | 0.77 | 0.82 | 2023 |
| Accuracy |  |  | 0.79 | 3187 |
| Macro avg | 0.77 | 0.79 | 0.78 | 3187 |
| Weighted avg | 0.80 | 0.79 | 0.79 | 3187 |

A blue squares with white squares

Description automatically generatedFIGURE 6.  A confusion matrix of the training dataset

A screenshot of a graph

Description automatically generatedFIGURE 7.  A confusion matrix of the testing dataset

A graph with blue and white bars

Description automatically generated with medium confidenceFIGURE 8.  A bar chart of feature importance (decreasing order)

VI. CONCLUSION

The

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES

*Basic format for reports:*

J. K. Author, “Title of report,” Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. *xxx*, year.

*Examples:*

1. E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.
2. J. H. Davis and J. R. Cogdell, “Calibration program for the 16-foot antenna,” Elect. Eng. Res. Lab., Univ. Texas, Austin, TX, USA, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

*Basic format for books (when available online):*

J. K. Author, “Title of chapter in the book,” in *Title of Published Book*, *x*th ed. City of Publisher, State, Country: Abbrev. of Publisher, year, ch.*x*, sec. *x*, pp. *xxx–xxx*. [Online]. Available: http://www.web.com

*Examples:*

1. G. O. Young, “Synthetic structure of industrial plastics,” in Plastics, vol. 3, Polymers of Hexadromicon, J. Peters, Ed., 2nd ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15-64. [Online]. Available: http://www.bookref.com.
2. *The Founders’ Constitution*, Philip B. Kurland and Ralph Lerner, eds., Chicago, IL, USA: Univ. Chicago Press, 1987. [Online]. Available: http://press-pubs.uchicago.edu/founders/
3. The Terahertz Wave eBook. ZOmega Terahertz Corp., 2014. [Online]. Available: http://dl.z-thz.com/eBook/zomega\_  
   ebook\_pdf\_1206\_sr.pdf. Accessed on: May 19, 2014.
4. Philip B. Kurland and Ralph Lerner, eds., *The Founders’ Constitution.* Chicago, IL, USA: Univ. of Chicago Press, 1987, Accessed on: Feb. 28, 2010, [Online] Available: http://press-pubs.uchicago.edu/founders/

*Basic format for journals (when available online):*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. *x*, no. *x*, pp. *xxx-xxx*, Abbrev. Month, year. Accessed on: Month, Day, year, doi: 10.1109.*XXX*.123456, [Online].

*Examples:*

1. J. S. Turner, “New directions in communications,” *IEEE J. Sel. Areas Commun*., vol. 13, no. 1, pp. 11-23, Jan. 1995. DOI. 10.1109.*XXX*.123456.
2. W. P. Risk, G. S. Kino, and H. J. Shaw, “Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle,” *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986, doi: 10.1109.*XXX*.123456.
3. P. Kopyt *et al., “*Electric properties of graphene-based conductive layers from DC up to terahertz range,” *IEEE THz Sci. Technol.,* to be published, doi: 10.1109/TTHZ.2016.2544142.

*Basic format for the most common types of unpublished references:*

a) J. K. Author, private communication, Abbrev. Month, year.

b) J. K. Author, “Title of paper,” unpublished.

c) J. K. Author, “Title of paper,” to be published.

*Examples:*

1. A. Harrison, private communication, May 1995.
2. B. Smith, “An approach to graphs of linear forms,” unpublished.