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

Yuta Sugiyama
The Department of Computational and Data Sciences
George Mason University
Fairfax, VA
ysugiyam@protonmail.com

Abstract— **Data breaches pose challenges for organizations, necessitating a proactive approach to mitigate risks and grasp the underlying dynamics. This study employs machine learning, data science, and Natural Language Processing (NLP) to predict Massachusetts data breaches, investigating organizational susceptibility, data protection strategies, and breach motivations. A Decision Tree classifier achieved 57% accuracy in training and 64% in testing, focusing on the majority classes of "Finance," "Healthcare," and "Other" organizations. The research underscores the necessity of robust data collection, enhanced NLP, and model refinement. Future work includes optimizing NLP categorization, exploring diverse machine learning models, fine-tuning hyperparameters, and collecting additional data for improved predictive accuracy, such as geographical locations. The study highlights the complex challenges posed by data breaches in diverse organizational contexts.**

Keywords— Natural Language Processing (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 occur 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. Consider how major retail chains have suffered significant financial losses and damage to their reputation due to data breaches. In September 2017, Equifax, one of the largest consumer credit reporting agencies, experienced a massive data breach compromising the sensitive personal information of 148 million Americans, including names, addresses, social security numbers, and credit card details [2]. The breach, due to a vulnerability in Equifax's online dispute portal, resulted in widespread concerns about identity theft and financial fraud, leading to a historic settlement offering affected consumers free credit monitoring, cash payments, and additional safeguards [2]. The breach resulted in significant financial losses for individuals, potential identity theft risks, and severe damage to Equifax's reputation, leading to extensive legal actions, congressional hearings, and ongoing investigations by federal and state authorities [2]. 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 explore the complex relationships between data breaches and their effects, equipping stakeholders with the tools and knowledge needed to prevent and mitigate these cyber threats effectively.

II. Data

The dataset chosen is from the state of Massachusetts (MA) Office of Consumer Affairs and Business Regulation website, holding reports of data breaches from 2007 to 2023 [1]. Originally in PDF format, the reports were consolidated into a single .csv file, featuring data from 2007 to 2022. The 2023 dataset was isolated for comparison with the predicted model.

The original dataset comprises 13 columns and 22,994 rows, totaling 1.5MB [1]. The "Assigned Breach Number" in the first column uniquely identifies each breach incident [1], while the second column, "Date reported," contains the date of the breach [1]. "Organization Name" in the third column provides categorical data on affected organizations [1], and the fourth column, "Breach Type Description," details the format of compromised data [1]. The fifth column, "Breach

Occurrence at Reporting Entity?" is a Boolean indicating if the breach occurred at the reporting entity [1]. The sixth column, "MA Residents Affected," contains numerical data on affected Massachusetts residents [1]. Columns 7-10 ("SSN Breached?", "Account Number Breached?", "Driver's Licenses Breached?", "Credit Debit Numbers Breached?") hold Boolean values indicating whether targeted data were compromised [1]. Columns 11-13 ("Provided Credit Monitoring," "Data Encrypted," "Mobile Device Lost Stolen") hold Boolean values indicating whether specific security measures were in place prior to the breach [1].

Assumptions were made within the dataset, which were inquired to the MA Office of Consumer Affairs and Business Regulation:

1. Were the missing columns within the Boolean columns within the dataset, count as the value "No"?
2. Do any of the other columns, besides the Boolean types, have any "False null data"?
3. In the column "Breach Type Description", there is a class called "undefined". Does that count as null data?

A response was given by the source, which confirmed these assumptions, which paved the way for the data cleaning process.

III. Data Cleaning

The data cleaning process included replacing null (NaN) values to "No" within all variables with boolean classes, dropping rows with "True null data" (any variables that does not hold boolean classes), and dropping the unnecessary column "Assigned Breach Number" as it was redundant to reaching our objective.

Outliers were identified in the original dataset through analysis of Table I. The maximum value of approximately 3,000,000 signifies instances of data breaches affecting 3,000,000 Massachusetts residents, with a mean of 715, representing the average number of residents affected. The high standard deviation of 24198 indicates significant dispersion of data points from the mean. This wide dataset range, attributed to numerous outliers, is visually evident in Figure 1.

TABLE I. STATISTICS OF THE NUMERICAL COLUMN

Column	MA Residents Affected
Minimum	0
Maximum	2982421
Mean	715
Median	3
Standard Deviation	24198

These outliers were detected by employing the Inter-Quartile Range (IQR) method. It was decided that deleting these rows would be appropriate to outlier handling due to the model objectives, which require the average amount of cases of data breaches. In Figure 2, a boxplot represents the data with the outliers removed. Again, a new statistic table was generated, giving more realistic results (Table II.)

TABLE II. STATISTICS OF THE NUMERICAL COLUMN (OUTLIER HANDLING)

Column	MA Residents Affected
Minimum	0
Maximum	41
Mean	12
Median	3
Standard Deviation	15

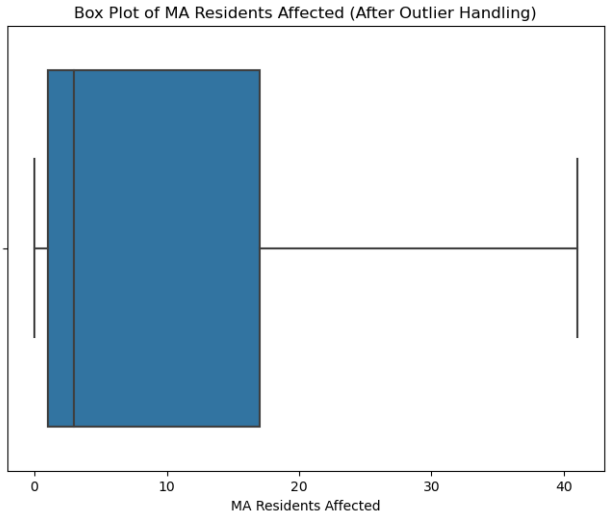

Fig. 1. A box plot of the originial dataset after to outlier handling

IV. Data Transformation And Feature Engineering

The data transformation process involved encoding Boolean columns, converting "Yes" and "No" to "True" or "False" values for model compatibility. Feature engineering utilized Natural Language Processing (NLP) on the "Organization Name" variable, creating "Organization Categories" with categorical values reflecting organization types. Challenges arose from ~8000 unique keywords, making it difficult to classify all the unique keywords to sort them into the keyword dictionaries. An attempt to automate using Google's JSON API was successful but deemed impractical due to budget constraints ($5 per 1,000 searches, totaling $200 for 15,000 searches). This presents an avenue for future exploration.

V. Exploratory Data Analysis

By implementing Exploratory Data Analysis (EDA), a facet plot, pie chart, and 3 bar charts were created to unravel and understand patterns, distributions, and relationships within the data chosen.

The facet plot in Figure 3 (A and B) depicts data breach types categorized by targeted data: driver's licenses, account numbers, credit/debit card numbers, and social security numbers (SSN). SSN stands out as the only targeted data with a significant number of undefined or electronic and paper breaches. Credit/debit card numbers have the highest count of electronic breaches, while driver's licenses have the lowest. Account numbers top the list for paper breaches, accounting for approximately 40% of total account number and SSN breaches.

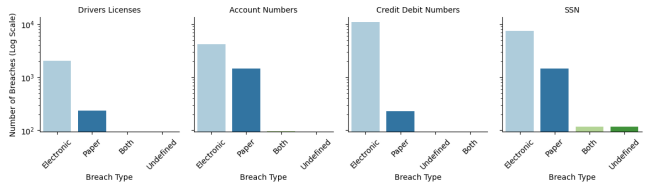

Fig. 3. A facet plot of the count of each data breach type by targeted data (y axis in log scale)

Figure 4's bar chart depicts data breach types from 2007 to 2022, with trendlines capturing changes over time. Electronic breaches dominate, showing consistent growth. A notable increase from 2010 to 2013 reflects technological shifts. Paper breaches rose steadily until 2016, declining afterward. Instances of combined electronic and paper breaches peaked in 2013 and gradually decreased. Cases with undefined breach types followed a similar pattern, peaking in 2013 and decreasing rapidly thereafter.

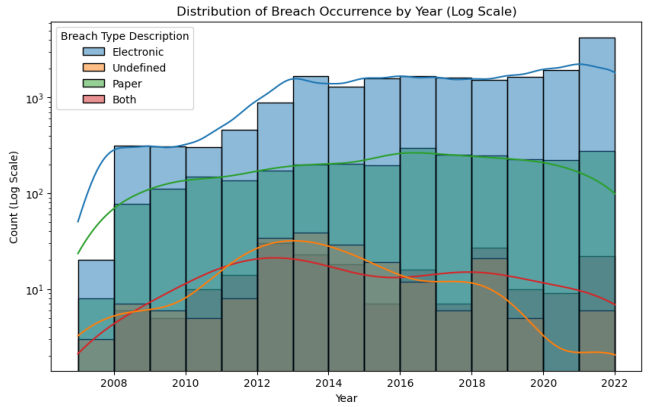

Fig. 4. A bar chart of the distribution of data breach occurences by year for each breach type (y axis in log scale)

Credit monitoring is a service, aiding individuals in monitoring and detecting suspicious credit-related activities. Offered by credit bureaus, financial institutions, or specialized companies, it helps users track changes to their credit reports. Figure 5's pie chart shows that during a data breach, 54.9% of organizations provided credit monitoring, while 45.1% did not, suggesting that just over half offered security services for monitoring credit transactions.

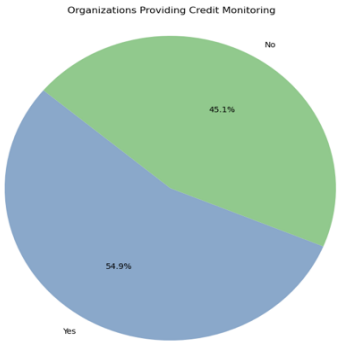

Fig. 5. A pie chart of the percentage of organizations with or without provided credit monitoring

Data encryption is the crucial process of securing information through coding to thwart unauthorized access during transmission or storage. It safeguards sensitive data, maintaining privacy and communication integrity. In Figure 6's bar chart, only 0.2% of breached organizations had encryption. This suggests encryption's potential importance in the model, though not insignificant.

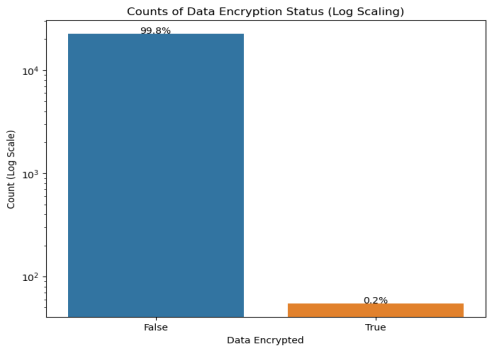

Fig. 6. A bar chart of organizations with or without data encryption at the time of the data breach (y axis in log scale)

Lost or stolen mobile devices in data breaches pose a significant security risk due to the potential exposure of sensitive information, including emails and corporate network access. Figure 7's bar chart shows that 2.8% of organizations experienced mobile device loss or theft during a breach, likely reflecting a period (2007-2009) with fewer security measures for mobile devices. Unauthorized access to critical data remains a concern, highlighting the importance of robust security practices.

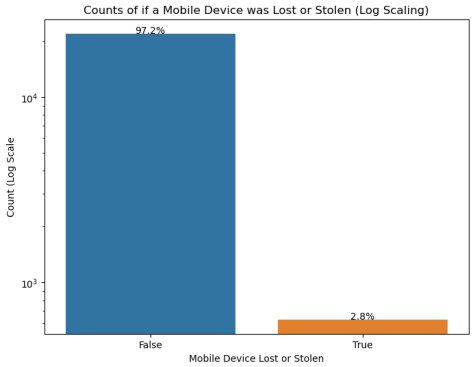

Fig. 7. A bar chart of organizations that had a mobile device lost or stolen at the time of the data breach (y axis in log scale)

VI. Model

The decision to employ a decision tree classification model for this project was driven by its effectiveness in achieving the three primary objectives. The model's capacity to capture complex, non-linear data relationships makes it adept at identifying patterns and decision rules crucial for predicting data breaches. The interpretability of decision trees adds transparency to the decision-making process, facilitating a better understanding and communication of insights regarding potential breach scenarios.

To address class imbalance, the RandomUnderSampler was utilized, randomly removing instances from the majority organization class. This ensures a more balanced distribution of classes during training, enhancing model performance and preventing bias towards the majority class, especially when the minority class (e.g., Healthcare) is of particular concern. Normalization was applied to scale data features consistently, preventing any single feature from dominating the model due to its scale. Denormalization post-training

allows interpretation and communication of results in the original dataset units, offering a more intuitive understanding of predictions within the context of specific data breach attributes.

The use of GridSearchCV facilitated the identification of optimal hyperparameters for the model. Hyperparameters such as max depth (set at 5), minimum samples at leaf nodes (set at 4), and minimum samples to split internal nodes (set at 2) were tuned to values that maximized performance on the training data. These optimization techniques, combined with a 75% training dataset and 25% test dataset split, contribute to a robust and well-performing model.

VII. Model Analysis And Evaluation

The model achieved an accuracy of approximately 57% on the training set in which it correctly predicted the class The model achieved 57% accuracy on the training set, correctly predicting class labels. In the "Finance" category, it showed 67% precision and 72% recall, resulting in a balanced F1-score of 69%. However, predicting healthcare organizations posed challenges with 52% precision, 28% recall, and an F1-score of 36%, indicating a trade-off between false positives and missed instances (Seen in Table III.)

Figure 8's training dataset confusion matrix breaks down the model's performance across Finance, Healthcare, and Other categories. Finance achieves a high 71.69% true positive rate, while Healthcare and Other show lower rates of 27.94% and 71.04%, respectively. Misclassifications, notably in Healthcare, are evident in off-diagonal elements, with 20.67% and 51.38% of instances wrongly predicted as Finance and Other. Healthcare predictions require improvement, addressing both false positives (10.89%) and false negatives (20.67%). Enhancing the model's ability in healthcare, possibly through targeted feature engineering or algorithmic adjustments, is crucial for accurate data breach identification.

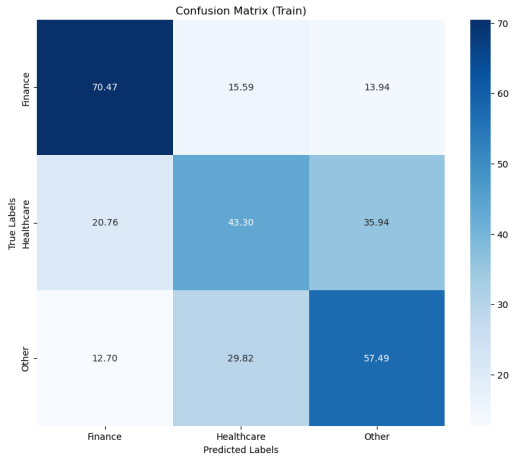

Fig. 8. A confusion matrix of the training dataset

TABLE III. Training Dataset

	Precision	Recall	F1-Score	Support
Finance	0.67	0.72	0.69	2462
Healthcare	0.52	0.28	0.36	2462
Other	0.51	0.71	0.59	2462
Accuracy			0.57	7386
Macro avg	0.57	0.57	0.55	7386
Weighted avg	0.57	0.57	0.55	7386

The model attained a 64% accuracy on the training set, correctly predicting class labels for a corresponding portion of instances in the testing data (see Table IV). Results were consistent between the training and testing datasets. Notably, for financial organizations, the model demonstrated high accuracy (80%) with balanced precision (70%) and recall, yielding a well-balanced F1-score of 75%. However, challenges arose in predicting healthcare organizations, with lower precision (27%), recall (28%), and an F1-score of 27%. Predictions for other organizations achieved a moderate level of accuracy (61%), balanced precision (70%), and a well-rounded F1-score of 65%.

Figure 9 displays the confusion matrix for the testing dataset. In predicting "Finance," the model shows a robust true positive rate of 70.26%, indicating its efficacy in correctly classifying instances under this category. However, challenges arise in predicting "Healthcare," with a lower true positive rate of 69.51% and notable false negatives (14.47%), highlighting the model's difficulty in accurately identifying healthcare-related breaches. For the "Other" category, the model achieves a 28.25% true positive rate, indicating moderate accuracy. The false positive rate for "Other" is 10.67%, signifying instances where the model incorrectly predicts breaches in this category. Enhancements are needed, particularly in accurately identifying healthcare-related data breaches, a crucial aspect for project objectives.

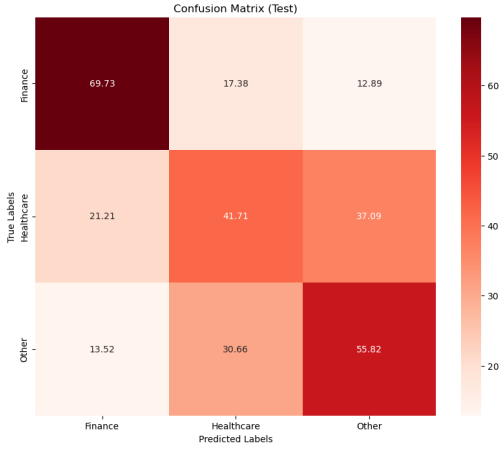

Fig. 9. A confusion matrix of the testing dataset

TABLE IV. Testing Dataset

	Precision	Recall	F1-Score	Support
Finance	0.80	0.70	0.75	2737
Healthcare	0.27	0.28	0.27	800
Other	0.61	0.70	0.65	2053
Accuracy			0.64	5590
Macro avg	0.56	0.56	0.56	5590
Weighted avg	0.65	0.64	0.64	5590

Figure 10's feature importance plot reveals key insights into predicting data breaches. "Credit Debit Numbers

Breached" is the most crucial feature, with a high 55.02% importance score (Table V). This underscores the significant impact of compromised credit and debit card numbers on the model's predictions. "Provided Credit Monitoring" is also influential at 21.63%, indicating its importance in shaping predictions. Other features like "MA Residents Affected" (9.21%), "Account Number Breached" (6.70%), and "SSNBreached" (4.86%) contribute, though to a lesser extent. Notably, breach type features (Electronic, Paper, Both) have minimal importance (0%), suggesting the limited impact of breach nature, likely due to the prevalence of electronic breaches.

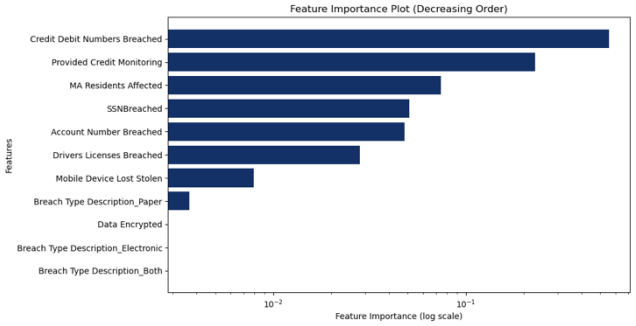

Fig. 10. A bar plot of feature importance (decreasing order)

TABLE V. FEATURE IMPORTANCE

Feature	Importance
Credit Debit Numbers Breached	0.55
Provided Credit Monitoring	0.22
MA Residents Affected	0.09
Account Number Breached	0.07
SSNBreached	0.05
Drivers Licenses Breached	0.02
Mobile Device Lost Stolen	0.003
Data Encrypted	0.002
Breach Type Description_Both	0.002
Breach Type Description_Paper	0
Breach Type Description_Electronic	0

The resulting decision tree, depicting a classification model predicting data breach types based on features. The root node sets the condition of "Credit Debit Numbers Breached" being ≤ 0.50. Breaches below this threshold are further categorized by SSN breaches, driver's license breaches, account number breaches, and the number of affected Massachusetts residents. Notably, breaches with credit and debit numbers below the threshold, SSN breached above the threshold, and a moderate number of Massachusetts residents affected are likely classified as "Other" incidents. Conversely, breaches with credit and debit numbers above the threshold, credit monitoring above the threshold, and a higher number of Massachusetts residents affected are predicted as "Healthcare" incidents.

A study tackled data breaches by creating an interactive visualization tool with the 2021 VAST Challenge Mini Challenge 2 dataset [4]. The tool aims to depict breach severity and deanonymization using user-friendly geospatial and transaction data visualizations [4]. The study emphasizes challenges in implementing the tool, such as issues with spatial data and the importance of careful project scope and data abstraction level considerations, highlighting the complexity of relations within data breaches [4].

VIII. CONCLUSION

This research embarked on the ambitious journey of predicting data breaches with a focus on data types and vulnerabilities using machine learning techniques, data science methodologies, many python-based data science libraries, and Natural Language Processing (NLP). The study aimed to reveal patterns and susceptibility factors in Massachusetts, offering insights into organizational vulnerabilities, data protection strategies, and breach motivations. Despite challenges like handling outliers and organization categorization complexity using NLP, the model achieved 57% accuracy in the training dataset and 64% in the test dataset. This emphasizes the potential of the model, especially in predicting breaches for major organizational categories such as Finance, Healthcare, and Other.

The model exhibited strong performance in identifying Finance-related incidents, as seen in the analysis of the train and test classification reports. Organizations should prioritize the protection of credit and debit card information, implementing robust security measures and encryption protocols to safeguard this sensitive data. Notably, the feature importance analysis emphasized the significance of variables like "Credit Debit Numbers Breached" and "Provided Credit Monitoring" in influencing breach predictions. Organizations should consider implementing credit monitoring services as part of their cybersecurity strategy, providing an additional layer of protection for individuals affected by a data breach.

The need for more practical approaches to categorizing organizations using NLP, downsizing the project for only the different types of financial institutions, and collecting supplementary data for improved accuracy has been identified as areas for future work.

It is evident that the study contributes valuable insights into the multifaceted dynamics of data breaches, providing a foundation for further refinement of predictive models and enhanced cybersecurity strategies.

REFERENCES

[1] Commonwealth of Massachusetts, “Data breach reports,” Mass.gov. Commonwealth of Massachusetts, 2023. Available: https://www.mass.gov/lists/data-breach-reports

[2] E. P. I. Center, “Equifax Data breach,” Electronic Privacy Information Center (EPIC), 2023. Available: https://archive.epic.org/privacy/data-breach/equifax/.

[3] “Custom Search JSON API,” Google for Developers. Available: https://developers.google.com/custom-search/v1/overview

[4] R. Carr, “Modified VAST Challenge with Applications to Data Breaches,” Modified VAST Challenge With Applications to Data Breaches, Available: https://www.cs.ubc.ca/~tmm/courses/547-22/projects/roz/report.pdfz