Understanding Data Breaches And Providing Support Using Data Science

Makena White, Annalia Schoenherr, Dr. Bukaita

College of Arts and Sciences, Lawrence Technological University

Introduction

The primary focus of this research is the impact of data breaches and the factors influencing them. Some industries are more vulnerable to these breaches compared to others. The increasing frequency of data breaches emphasizes the need to understand the underlying reasons. Apart from impacting millions of people and causing significant monetary losses, data breaches also impact factors like consumer trust and security. The goal as Computer Science students, is to investigate the factors and patterns contributing to the rise in data breaches, and forecast the possible data breaches to come. A significant example of one of the largest data breaches in history was the Equifax Data Breach in 2017, which exposed sensitive data of 147 million individuals, representing 40% of the US population. The statistics of this breach can be seen in Figure 1.

Although data breaches have occurred in the past, the most significant ones can be traced back to 2005 when multiple large-scale breaches were reported and tracked by the Privacy Rights Clearinghouse. These breaches were possible due to record digitization, which made record use and management easier but also made these systems more vulnerable to attacks. It is crucial to protect this data from malicious parties to prevent financial loss and emotional distress. Individuals often have to pay financial costs and cope with distress, such as stolen data like credit cards or medical records, if their sensitive information gets breached. Entities also risk losing customer trust and facing financial damages. Government entities face the risk of exposing their data to enemies of the state. Overall, data breaches are a crucial issue that needs to be addressed.

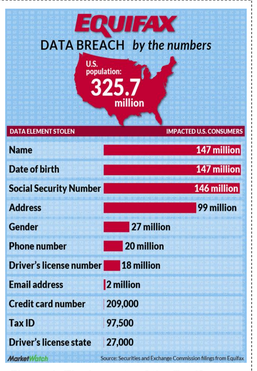


Figure 1- The impacts of the Equifax data breach of 2017

Statistical Data

The data used for analysis was collected from a research website called Comparitech. Researchers at Comparitech collect and analyze data breaches and incidents around the globe for public education. This dataset consists of records of data breaches. The sheet includes the name of the entity who experienced a data breach, they type of company they are, the number of individuals affected by the data breach, the breach submission date, the type of data breach, the location of the breached information, whether or not a business associate was present, and a description of the incident. From this data, the following graphs were extracted to analyze the frequency of each data breach occurrence. This data can be considered a discrete time series because the data is collected at specific intervals. Certain statistics can be used to analyze this data, including the mean, the variance, and correlation of each dataset. Each statistic can be seen as follows:

Mean Data Breach Statistics:
Hacking Incident Reports (per month): **2.69**
Improper Disposal Reports (per month): **0.67**
Loss Reports (per month): **1.16**
Theft Reports (per month): **8.60**
Unauthorized Access/Disclosure Reports (per month): **4.54**

Number of People Affected By Hacking: **1,383,735.89**
Number of People Affected By Improper Disposal: **20,541.9**
Number of People Affected By Loss: **105,153.83**
Number of People Affected By Theft: **273,853.82**
Number of People Affected By Unauthorized Access/Disclosure: **163,714.98**

Mean Standard Deviation:
Hacking Reports (per month): **0.265**
Improper Disposal Reports (per month): **0.079**
Loss Reports (per month): **0.133**
Theft Reports (per month): **0.296**
Unauthorized Access/Disclosure Reports (per month): **0.357**
Number of People Affected By Hacking: **269,441.27**
Number of People Affected By Improper Disposal: **2,573.99**
Number of People Affected By Loss: **19,095.71**
Number of People Affected By Theft: **36,997.74**
Number of People Affected By Unauthorized Access/Disclosure: **24,066.44**

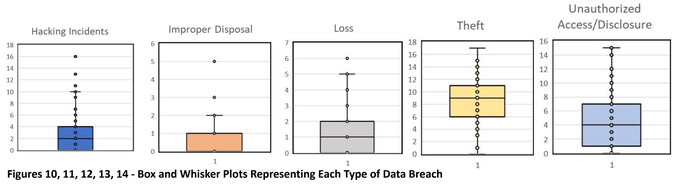
The mean value is used to provide a general insight into the dataset. To do this, the sum of each type of data breach was compiled and divided by the total number of months. This was completed for the number of occurrences and for the number of people affected. The results highlight multiple insights into the data. It can be determined that data breaches occur most commonly via theft, and least commonly via improper disposal. It can also be seen that hacking has the largest average impact on people, and that improper disposal has the smallest average impact on people.
Standard deviation measures how the data is dispersed in relation to the mean. It is calculated using the following formula:

$$\sigma=\sqrt{(\Sigma(\chi_i-\mu)^2)\ /\ N}$$

The mean value for standard deviation was found for each category, and the results highlight how little the values vary for data breaches resulting from improper disposal and how greatly the number of people affected by data breaches via hacking varies from its average.

Figure 2 - Frequency of Data Breach Occurrences Filtered By Type

Box-and-whisker plots are also used to graphically summarize data. The colored boxes in the following graphs represent the middle 50% of the data, with the top representing the upper quartile and the bottom representing the lower quartile. The points outside of each box on the line show the range of the data. The points not on the lines, represent outliers.



Figures 10, 11, 12, 13, 14 - Box and Whisker Plots Representing Each Type of Data Breach

Model Analysis

To make predictions about future data breaches, the data underwent multiple calculations, beginning with exponential smoothing.
Exponential smoothing is a forecasting technique that is used on data that often contains random data fluctuations. Exponential smoothing uses a weighting factor, represented by "α", that is assigned at a decreasing rate to the data. This method was chosen because there is no clear trend or seasonal pattern present in the data. The following formula is used for this process:

Exponential Smoothing Formula: $\hat{Y}_{t+1}=\hat{Y}_t+a(Y_t-\hat{Y}_t)$

Example Calculation: $\hat{Y}_{t+1}= 1230.26 + 0.35(2100-1230)$

In this formula, $\hat{y}_{t+1}$ represents the current forecast value, $\hat{y}_t$ represents the previous forecast value, and α represents the weighing factor. The factor chosen for this dataset was 0.35. A higher value of alpha would place more weight on the more recent data points, while a lower value for alpha would give more weight to older points in the dataset. This calculation was performed on all records occurring between October of 2009 and October of 2016. For the values being forecasted between November 2016 and December of 2024, an increment value had to be calculated for trend projection. The formula can be seen as follows:

Increment Formula = $(\text{Last } \hat{Y}_t - \text{First } \hat{Y}_t) / (n-1)$

Example Calculation (Hacking): $(35134.72 - 40) /(86-1)$

In this formula, the first value predicted using the exponential smoothing method is subtracted from the last value predicted, and the result is then divided by total number of predictions minus one. The dataset before forecasting included 86 rows of data for each type of breach. Examples of the value to increment by can be seen below:

Hacking	Improper Disposal	Loss	Theft	Unauthorized Access/Disclosure
408.08	5.52	30.30	79.39	47.21

To predict the next value for each data breach, the calculated increment value was added to the previously predicted value. This process was repeated until the desired forecast date was reached.

To ensure the data was predicted with accuracy, the mean absolute deviation was calculated. The MAD measures the average absolute difference between the actual value and the predicted value. The result indicates how well the model is able to predict the data with a result closer to 0 indicating greater accuracy. To do this, the absolute difference between the actual value and the predicted value is calculated for each time period. The differences are then averaged to determine the overall mean absolute deviation for the data. An illustration of the MAD for the frequency of data breaches by type can be seen as follows:

Hacking	Improper Disposal	Loss	Theft	Unauthorized Access/Disclosure
1,378,783.52	18,884.14	103,446.71	268,291.18	159,597.17

Once these steps were completed, the forecasted data was produced as seen in figures 3 - 8. Note that the original data lies before the dotted red line, and the forecasted data lies after the dotted red line in each graph. Each graph includes the recorded/forecasted month on the x axis, and the number of occurrences on the y axis.

An additional method of analyzing the data is to determine if there is a correlation that exists between two factors. To analyze correlation, a data manipulation and visualization tool called "RStudio" can be used. The data seen in Figure 9 denotes the correlation between each of the data factors including month and type of data breach. Correlation is an important tool because it shows if certain variables impact each other.

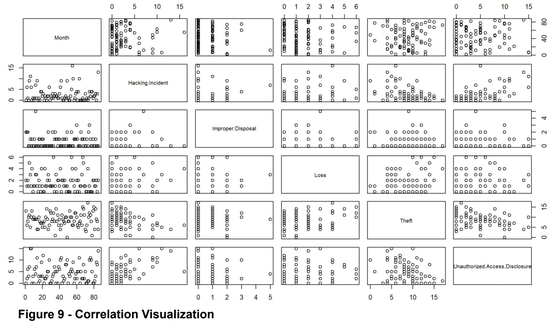


Figure 9 - Correlation Visualization

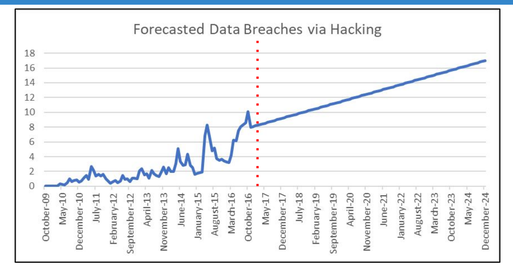


Figure 3 - Forecasted Data Breaches via Hacking
This chart shows the number of data breach occurrences that resulted from hacking. The forecasted data increases over time, indicating that future data breaches resulting from hacking could increase.

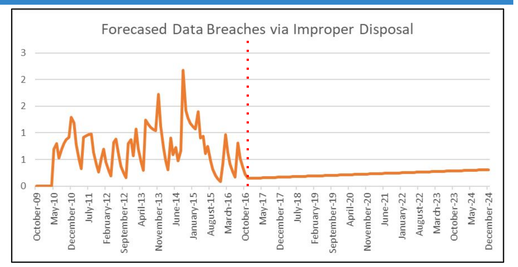


Figure 4 - Forecasted Data Breaches via Improper Disposal
This chart shows the forecasted prediction of data breach occurrences resulting from improper disposal of personal data by an entity. The data fluctuates with no clear trend, and the predicted forecast shows low and steady occurrences.

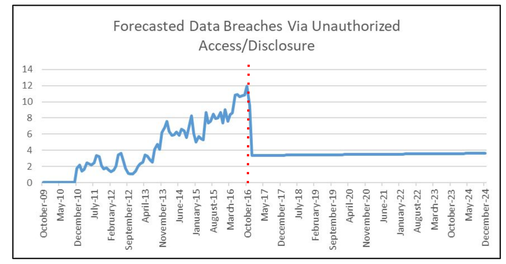


Figure 5 - Forecasted Data Breaches via Unauthorized Access/Disclosure
This chart shows the forecasted prediction of data breach occurrences resulting from unauthorized access or disclosure by an entity.

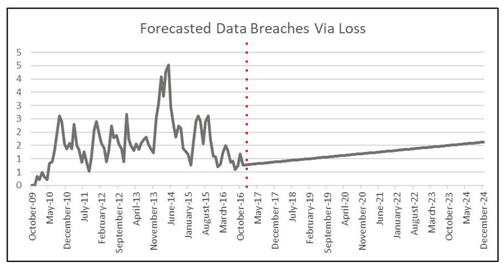


Figure 6 - Forecasted Data Breaches via Loss
This chart shows the forecasted prediction of data breach occurrences resulting from an entity losing personal data.

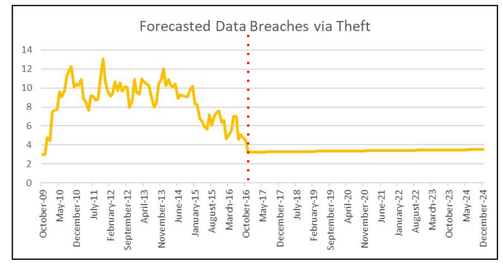


Figure 7 - Forecasted Data Breaches via Theft
This chart shows the forecasted prediction of data breach occurrences resulting from theft from an entity.

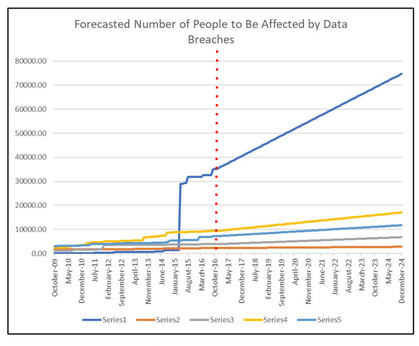


Figure 8 - Forecasted Number of People Affected By Data Breaches
This chart shows the forecasted prediction of the number of people who could be affected by a data breach. Each line denotes a method of breaching data- including hacking, improper disposal, unauthorized access/disclosure, loss, and theft.

Conclusion

The main objective of this project was to investigate how data breaches are becoming more prevalent and which factors may continue to occur in the future. By analyzing data breaches, the group identified key factors in understanding and predicting potential vulnerabilities when it comes to data breaches. Based on the data analysis and forecasting, the prediction is an upward trend in data breaches overall as well as the number of people affected by data breaches. Another prediction is industries with highly sensitive data, such as healthcare and finance, will continue to be susceptible to data breaches. The data breach method that was found to increase the most is hacking (See Figure 2). Due to the world becoming more digitized, more and more personal data becomes digitized as well. As seen based on the data analysis, the industries with highly sensitive data, such as healthcare and finance, will be large targets for hackers. Another possible reason for this is the advancement of hacking strategies. As technology advances, so do hacking methods. From social engineering to ransomware attacks, hackers are constantly developing new ways of infiltrating companies' systems. Lastly, another potential reason for the upward trend of data breaches is human error. Factors such as weak passwords, inadequate handling of sensitive information, and theft.

The purpose of the data analysis is to explain why some industries are more susceptible to data breaches. Utilizing data science can help to identify factors to understanding data breaches. Overall, the aim of this data analysis is to find ways to prevent these detrimental breaches.

Some ways to better protect against data breaches include educating employees on the risks of social engineering, incorporating better security practices, such as two-factor authentication or more secure encryption, and informing users of the importance of cybersecurity, such as choosing secure passwords.

These solutions will hopefully provide a safer digital future. By utilizing data analytics, companies can better protections against data breaches as well as limit their impact.

References

Epic - Equifax Data Breach. Electronic Privacy Information Center. (n.d.). https://archive.epic.org/privacy/data-breach/equifax/

Moody, R. (2023, January 6). *Map of worldwide ransomware attacks (updated daily)*. Comparitech. https://www.comparitech.com/blog/information-security/global-ransomware-attacks/

Vigderman, A. (2023, August 11). *What is a data breach and how to prevent a breach in 2024*. Security.org. https://www.security.org/identity-theft/what-is-a-data-breach/#biggest