Predicting Data Breaches in the Healthcare Industry

Final Report

Mandi Uppal

1. Introduction:

The cost of poor cybersecurity practices is paramount. In particular, the healthcare industry experiences the highest number of attacks by ransomware, in addition attacks are expected to quadruple by 2020 (1). Currently 41 percent of companies have over 1,000 sensitive files including credit card numbers and health records left unprotected (2). Financial costs can look like billions of dollars as was the case with the WannaCry virus of 2017 that affected 100,000 groups in hundreds of countries and more than 400,000 machines that were infected (3). It's important to note that organizations that invest money on preventing cybersecurity can expect to only spend a small fraction of the expected loss (4).

As a consultant for a Canadian healthcare startup that is establishing their cybersecurity programs and want to ensure that their clients are protected from potential data breaches, my job is informing preventative cybersecurity practices and procedures. I will determine the risk level of data breaches occurring and identify factors that may contribute to breach occuring

I will use the U.S. Department of Health and Human Services Office for Civil Rights data breaches dataset to determine the level of risk this startup is vulnerable to by comparing companies and states where data breaches are occurring and common patterns among these breaches such as the number of individuals affected by data breaches among other factors. The dataset that will be used in this project is listed below:

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

I'll be using machine learning to make predictions on the drivers behind data breaches. Deliverables for this project will include code and a report on my methods and findings on what drivers are behind data breaches and areas that the company may be vulnerable based on the patterns I have discovered.

2. Dataset:

The dataset being used in this project is called the "U.S. Department of Health and Human Services Office for Civil Rights Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information". This dataset was obtained online from the U.S. Department of Health and Human Services Office for Civil services website on August 26, 2019, and contains archives of all resolved breach reports older than 24 months.

The dataset contains the following features: “Name of Covered Entity”, “State”, “Covered Entity Type”, “Individuals Affected”, “Breach Submission Date”, “Type of Breach”, “Location of Breached Information”, “Business Associate Present”, and “Web Description”.

3. Data Cleaning and Data Wrangling:

I. Purpose:

The aim of this capstone project is to discover predictive drivers of data breaches within the healthcare industry. The questions guiding the project include what the risk level of any given company in the healthcare industry experiencing a data breach is and what type of breach is most likely .

II. Steps:

i. Datetime

The US HH dataset was downloaded from the https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf website from the archive tab as a csv file and imported into jupyter notebooks as a dataframe in python. Using the df.head() method the structure of the dataframe was inspected noting column names. Next, data types were explored using df.dtypes in order to determine if any format changes were needed. After inspection of the data types of column values it was determined that the “Breach Submission Date” was stored as an object type and would need to be transformed into the type datetime for efficient analysis, this was accomplished but applying the code pd.to_datetime(df[‘Breach Submission Date’]). After this transformation and asserting that the transformation was successful using print(df.info()) to view column types, the size of the dataframe was ascertained using df.shape, which returned 2315 rows and 9 columns.

ii. Missing Values

Next, all columns were checked for null values using df.isnull() and chaining .sum() to the method in order to return the cumulative nulls per column. Noting the columns with null values, a new dataframe called “dropped” was created by applying df.dropna() in order to drop all columns with null values. Using dropped.shape the “dropped” dataframe was inspected to confirm that the 232 null rows were removed from the dataframe, the results of the shape method confirm that 2083 rows remain in the in “dropped” dataframe.

iii. Duplicate Values

After dropping null values the next step in the data cleaning process was to determine if duplicate values occurred in columns. This was determined by applying unique = len(dropped['Name of Covered Entity'].unique()) on each column. Note, this method returns the number of rows in column that are unique. For the column “Name of Entity Covered” we can see that of the 2083 rows in the “dropped” dataframe only 1899 are unique meaning that further inspection is needed in order to determine why duplicates are occuring. Where duplicates were occuring was examined by applying df_2 = dropped.groupby(['Name of Covered Entity']).count() in order to return the result of number of times every company name occurs. This was further examined by filtering where company names occurred more than once using df_2 = df_2[df_2['State'] > 1]. Further investigation shows that for example the company “Clearpoint Design, Inc.” appears multiple times as a result of experiencing multiple data breaches of different dates. This was determined by applying df_3 = dropped[dropped['Name of Covered Entity'] == 'Clearpoint Design, Inc.'] in order to see all instances of where “Clearpoint Design, Inc.” occurs. The above steps were applied to the column state to discover that there are 52 states listed instead of the expected 50. Upon manual inspection of the result it was determined that the extra two state values listed were ‘PR’ and ‘DC’. Using print(dropped['Covered Entity Type'].unique()) returns 4 unique categories for “Entity Type” (Business Associate, Health Plan, Healthcare Provider, Healthcare Clearing House) represented in the data. Unique values were assessed using the steps above for the columns name of entity covered, state, covered entity type, type of breach, location *of breached information.*

iv. Names

The column “Name of Entity Covered” did not have uniform formatting, using regular expressions this was corrected. The following symbols were removed, as were the words using the code:

```
dropped['Name of Covered Entity clean'] = dropped.loc[:,'Name of Covered
Entity'].replace(r'[/|\.|\"|\,|\&|\()|\[]|\-|\'|\]', "",
regex=True).str.replace(r'\binc\b|\bllp\b|\bcorp\b|\bllc\b|\bpllc\b|\bpc\b|\bdba\b|\band\b',
"", case=False, regex=True)
```

The words and symbols in .replace() are removed from the column values. The result returns fields without duplicate company names due to formatting variances.

v. Binary Categories for Breach Types

The values of ‘Type of Breach’ column were separated into unique values in order to sum unique values for further analysis as values in this column were grouped into multiple categories. Separation into unique value columns in dataframe was achieved by creating a list assigning all unique values with the following code: type_of_breaches = ['Hacking/IT Incident', 'Other', 'Unauthorized Access/Disclosure', 'Theft', 'Improper Disposal', 'Loss', 'Unknown']. New columns were created and added to the dataframe by iterating though the list of unique values using a for loop and assigning each value a 1 or 0 for where it occurs as a value in the dataframe. This was accomplished using the code:

```
for value in type_of_breaches:

        dropped[value] = np.where(dropped['Type of Breach'].str.contains(value), 1, 0)
```

The sum of these new columns is calculated by applying np.sum() on the dataframe and unique value column name as follows: np.sum(dropped['Hacking/IT Incident']).

4. Data Storytelling:

I. ***Visual EDA***
II. ***Numerical EDA***

In this exploratory data analysis, the problem I am exploring is the factors affecting data breaches. I investigated various questions to provide insights to this problem. Detailed in this report is the approach used accompanied by my findings.

i. Top Breach Types:

As a preliminary step I counted the frequency of various variables. First, I counted the number of unique breach types and plotted as a graph, the result shows that the breach category of “Theft” is the most common breach type followed closely by “Unauthorized Access/Disclosure”, and “Hacking/IT Incident” composing the top three common breach types.

ii. Top Ten Breach Locations:

Next I explored the top ten breach locations, the results show that “Paper/Films” is the top offender followed closely by cybersecurity related categories: “Network server”, “laptop”, and “email”. Upon further investigation of the location of breached information grouped by only cybersecurity categories (“Hacking/IT”) we can see that data breaches via “Desktop Computer” are most frequent.

iii. Top Ten States with the most breaches

Additionally I also counted the the top ten states with the most breaches which were discovered to be “California”, “Texas”, and “Florida”.

iv. Number of Individuals Affected by Breach Type:

After uncovering the top three common breach types (“Unauthorized Access/Disclosure”, and “Hacking/IT Incident”), this lead me to ask the questions, “What is the effect of these breach types**?”** and “How impactful are these “top” breach types on the number of individuals affected**?”.** To answer these questions I calculated the total number of individuals affected by type of breach and found that “Hacking/IT Incident” type breaches impacted the most number of individuals, followed by “Theft”, “Loss”, and “Unauthorized Access/Disclosure” to lesser extents.

Further investigation into the number of individuals affected by unique breach types did not unveil any pattern when plotted as individual scatter plots of each breach type and individuals affected (Figure 1).

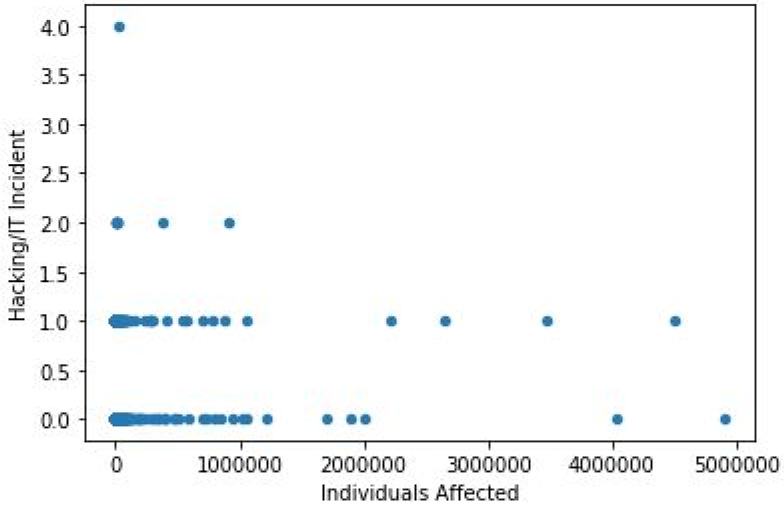


Figure 1: Number of Individuals affected by “Hacking/ IT Incident” breach type. The x-axis begins at 500 individuals.

v. Companies with the most breaches:

Following individuals affected I explored the question, “What ten companies have the most breaches?”. The ten companies with multiple breaches occurring the most are displayed as a bar chart of frequency of breaches by company in Figure 2 below:

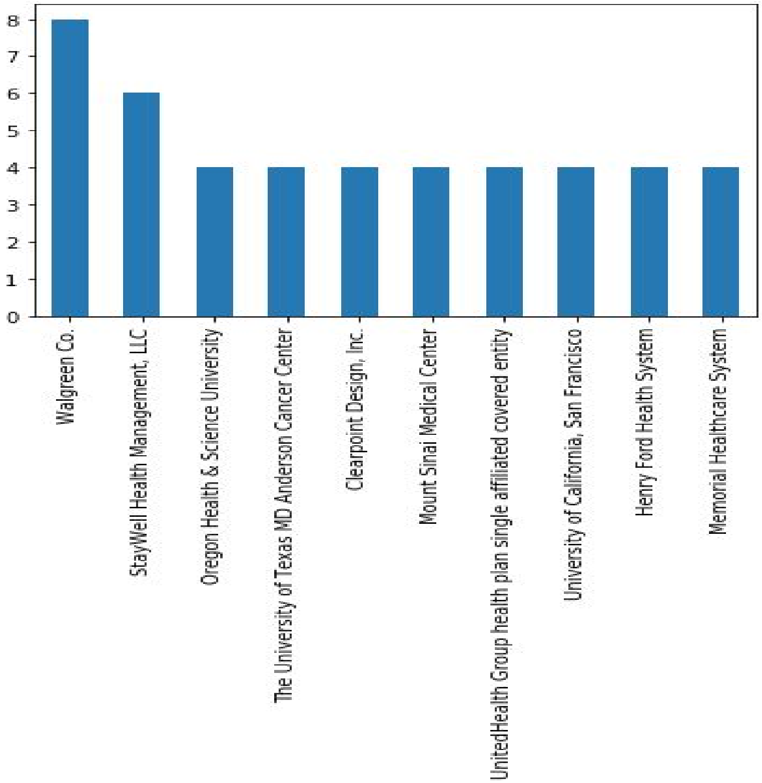


Figure 2: Top ten companies with multiple breaches.

vi. Multiple breaches within the same company

I then looked at companies where the same breach type occurred more than once. This exploratory data analysis shows that one company had 4 instances of “Hacking/IT Incident” breaches, while the other 8 companies with multiple “Hacking/IT Incident” breach types had 2 instances of the breach type. The types of breach affecting the most number of companies more than once are “Unauthorized Access/Disclosure”, and “Theft”. At first look it seems that “Hacking/IT incidents” are unlikely to occur more than once at a company. However this does not mean that cybersecurity data breaches are less likely to occur, further exploration is needed. Additionally, this insight does not speak to the impact of one instance of a cybersecurity related data breach on the number of individuals affected.

vii. Breaches over time:

In order to get a picture of what the trend is in data breaches over time is I plotted a time series of companies affected over time (Figure 3). The results indicate that the frequency of data breaches increased over time, peaking in 2014 and has generally decreased over time since then.

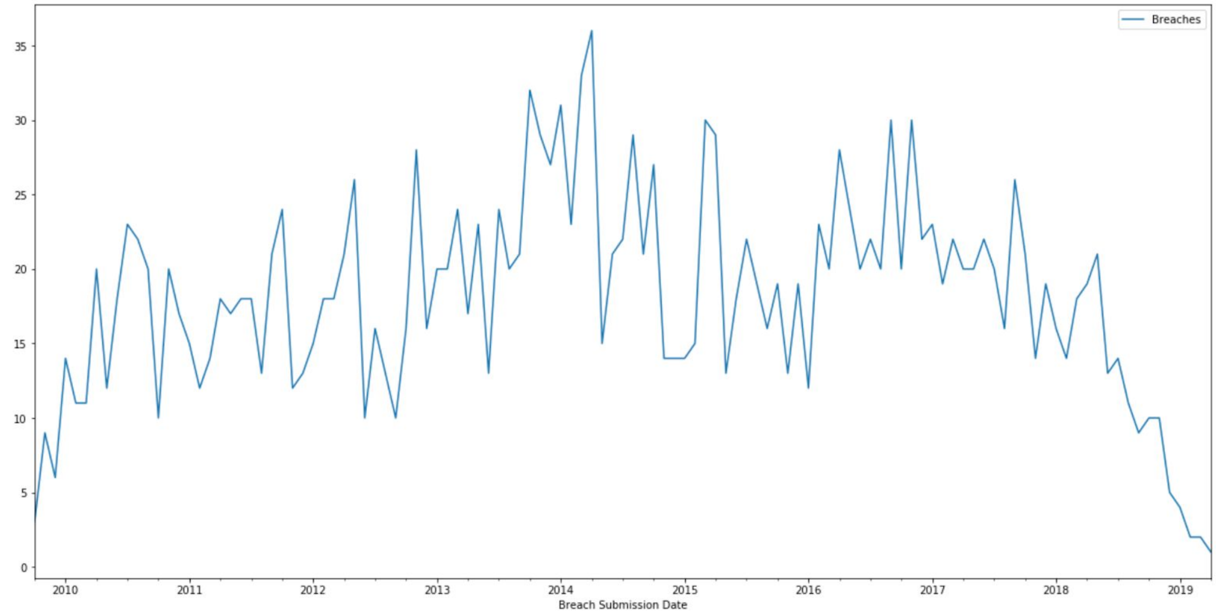


Figure 3: Data breaches over time.

III. Statistical Analysis

This section summarizes the statistical analysis performed on the number of individuals affected by “Hacking/IT Incident” type breaches versus other breaches, whether or not there is a relationship between “Hacking/IT incident” type breaches and the entity type affected, and whether the number of individuals affected by hacking vs other breach and entity type.

Variables

I. Individuals affected by cybersecurity breach types

i. T-test hypothesis test

The null hypothesis that the number of individuals affected by "Hacking/IT Incident" breach types are the same as the number of individuals affected by breach types that are not "Hacking/IT Incident" breach types and the alternative hypothesis that the number of individuals affected by "Hacking/IT Incident" breach types are not the same as the number of individuals affected by breach types that are not "Hacking/IT Incident" breach types were examined using a t-test hypothesis test.

The resulting p-value was 0.0000001638223020546479 therefore we reject the null hypothesis, in other words there is a difference between the number of individuals affected by hacking/it breaches and other breach types. We can therefore utilize this variable relationship when building a machine learning model.

II. Categorical variables:

i. Entity types affected by cybersecurity breach types

The null hypothesis that there is no difference between “Hacking/IT Incident” breach types and other breach types across entity types and the alternative hypothesis that there is a difference between “Hacking/IT Incident” breach types and other breach types across entity types was examined using a chi squared test.

Using a chi square test analysis, the resulting p-value was 0.11 which is less than the alpha level of 0.05 and therefore we fail to reject the null hypothesis. In other words there is no difference in “Hacking/IT Incident” breach types and other breach types across entity types.

ii. Individuals affected by breach and entity type

The null hypothesis that there is no difference between breach type (Hacking/IT Incident vs. other) and entity type over individuals affected and the alternative hypothesis that there is a difference between breach type (Hacking/IT Incident vs. other) and entity type over individuals affected was examined using a chi square test.

Using a chi square test, the resulting p-value was zero which is less than alpha of 0.05, and therefore the null hypothesis is rejected null. In other words, there is a difference in individuals affected by “Hacking/IT incident” type breaches and entity type.

However, it is uncertain whether the resulting difference is due simply to the known difference we discovered during previous analysis between the number of individuals affected and breach type. Further analysis could examine additional categorical columns to see if the difference persists, if the same pattern continues it may simply be due to the relationship between breach type and individuals affected.

iii. Location of breach methods

The null hypothesis that there is no relationship between type of breach and location of breach and the alternative hypothesis that there is a relationship between type of breach and location of breach was examined using a chi-square test.

Using a chi square test, the resulting p-value was 1.267e-13, which is less than alpha of 0.05, and therefore the null hypothesis is rejected null. In other words, there is a relationship between type of breach and location of breach. We can use these variables to inform our machine learning model.

5. Conclusion and Next Steps:

Summary

Findings

We have examined various variables in order to attempt to gain a better understanding of which ones are meaningful in predicting future data breaches. We see clear relationships for location of breach method, number of individuals affected and the type of breach.

Next Steps

The next part of this project will include building a machine learning model using Random Forest, Naive Bayes, Decision Tree, and Logistic Regression.

Mandi Uppal

Predicting Cybersecurity Data Breaches in the Healthcare Industry: In Depth Analysis

Introduction

The problem examined in this project is predicting the occurrence of cybersecurity type data breaches within the industry of healthcare. These findings will in turn guide the development of prevention plans to protect against future cybersecurity data breaches from occurring. Modelling will validate the features most important in considering when creating prevention plans.

Machine Learning Models: *Deep Dive*

Data Preprocessing

In order to prepare the dataset to be used in a machine learning model, the following categorical columns were split into binary variables: “Location of Breach Method” and “Business Associate Present”. Additionally, the column “Breach Submission Date”, was split into three columns: day, month, and year. Additional features extracted from date were: the quarter, the day of the week, whether or not a date was a weekend, and the week of the year. NaN values were removed, and the dataset was scaled to exclude outliers, which in this case were observations of individuals affected greater than ten thousand. Due to the presence of greater incidences of other breach types (1412) compared to cybersecurity related breach types (256), classes were balanced using dataframe filtering. Arrays of equal observations for each class (other breach type and hacking/it incident breach type) were created to feed into machine learning models.

Models

In order to solve this classification problem of predicting breach type, the following models were applied to the dataset: Random Forest, Naive Bayes, Decision Tree, and Logistic Regression.

Metrics: *Accuracy, Precision, and Recall*

Model	**Accuracy**	**Recall**	**Precision**	**Model Rank**
Random Forest	0.85	0.80	0.85	2
Naive Bayes	0.74	0.95	0.64	4
Decision Tree	0.83	0.75	0.70	3
Logistic Regression	0.88			1

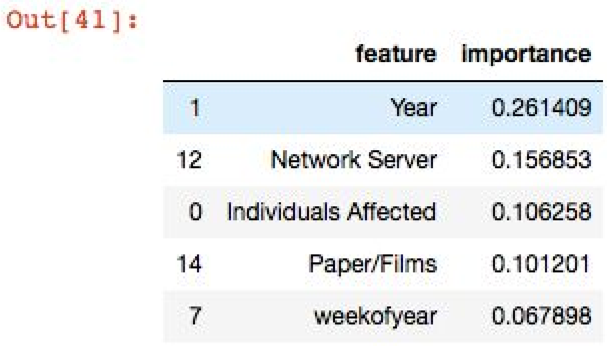
Figure 1: *Performance of models*

The results indicate that logistic regression and random forest classifiers were the best performers in this analysis. The Naive bayes model results indicate a poor model merely guessing breach type.

Hyperparameter Tuning

Random Forest

Using the feature importance variable, feature importance scores were determined and visualized.



Out[41]:

	feature	importance
1	Year	0.261409
12	Network Server	0.156853
0	Individuals Affected	0.106258
14	Paper/Films	0.101201
7	weekofyear	0.067898

Figure 2: *Feature importance scores for random forest model.*

A new model was generated model on important features, less important features were removed and only the top five important features are used including *year, network server, individuals affected, Paper/Films, and week of year*. Using only important features and setting n_estimators to 100, the model had a 5% increase in accuracy.

Logistic Regression

In order to find the best model for the logistic regression model, the regularization parameter *C* was tuned and the *GridSearchCV* function was used to perform cross validation and grid search over the training data. Implementation of the *GridSearchCV* function using 5 fold cross validation and using *C* parameters 0.0001, 0.001, 0.1, 1, 10, 100 indicated that the parameter C=10 performs the best in this model. The accuracy performance of this model on the test set is 0.81 compared to 0.77 for the initial model prior to the turning of the regularization parameter *C*.

	Accuracy
Model performance prior to regularization	0.77
Model performance after tuning C	0.81

Figure 3: *Performance of logistic regression models before and after tuning regularization hyperparameter C.*

Conclusion: Insights and Further Research

Further investigation is needed into commonalities of weeks important in prediction model. Year was another important feature as indicated by the random forest model, research further what was happening in these years? The number of individuals affected was another important features in predicting data breach type, more data is needed to see if there is a relationship between company size and breach type. Interesting that paper/films has importance, maybe there is less breach type overlap in this category. The data used in this project presents limitations due the importance of the location of breach methods as a feature. Try removing location based features: network server, and then paper/films. Removing 'Network Server' reduces model accuracy to 75% percent. What other factors are important after this removal? Removing both 'Network Server' and 'Paper/Films' reduces performance to 72.8 percent. Including 'Network Server' Increases to 84%, but how well will this generalize to new data aka new companies? Enhancing data to include more features will guide more meaningful results.

Sources:

1. https://www.csoonline.com/article/3237674/ransomware/ransomware-damage-costs-predicted-to-hit-115b-by-2019.html
2. https://info.varonis.com/hubfs/2018%20Varonis%20Global%20Data%20Risk%20Report.pdf
3. https://www.varonis.com/blog/cybersecurity-statistics/
4. https://en.wikipedia.org/wiki/Computer_security#cite_note-91