Predicting Data Breaches in the Healthcare Industry

Final Project Report
By: Mandi Uppal

Cybersecurity and Data Breaches

- 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).

Proposal

- 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 use the U.S. Department of Health and Human Services Office for Civil Rights Breach Portal data 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.
- I'll be using machine learning to make predictions on the drivers behind data breaches. Deliverables for this project include code and a report on methods and findings on what drivers are behind data breaches.

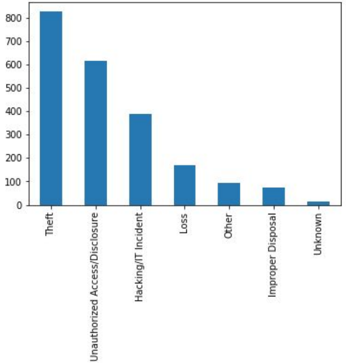
Data Wrangling

- Drop columns
  - Web Description
- Duplicates
- Missing Values
- Data Types
  - Breach Submission Date
- Clean fields
  - Name of Covered Entity
- Binary columns for categorical variables

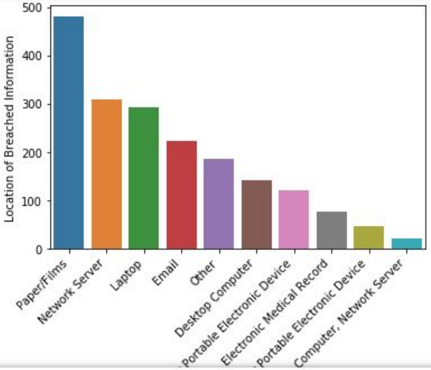
Exploratory Data Analysis

- Top breach types
- Top ten breach locations
- Top ten states with the most breaches
- Number of individuals affected by breach type
- Companies with the most breaches
- Breaches over time

Top Breach Type and Location of Breach

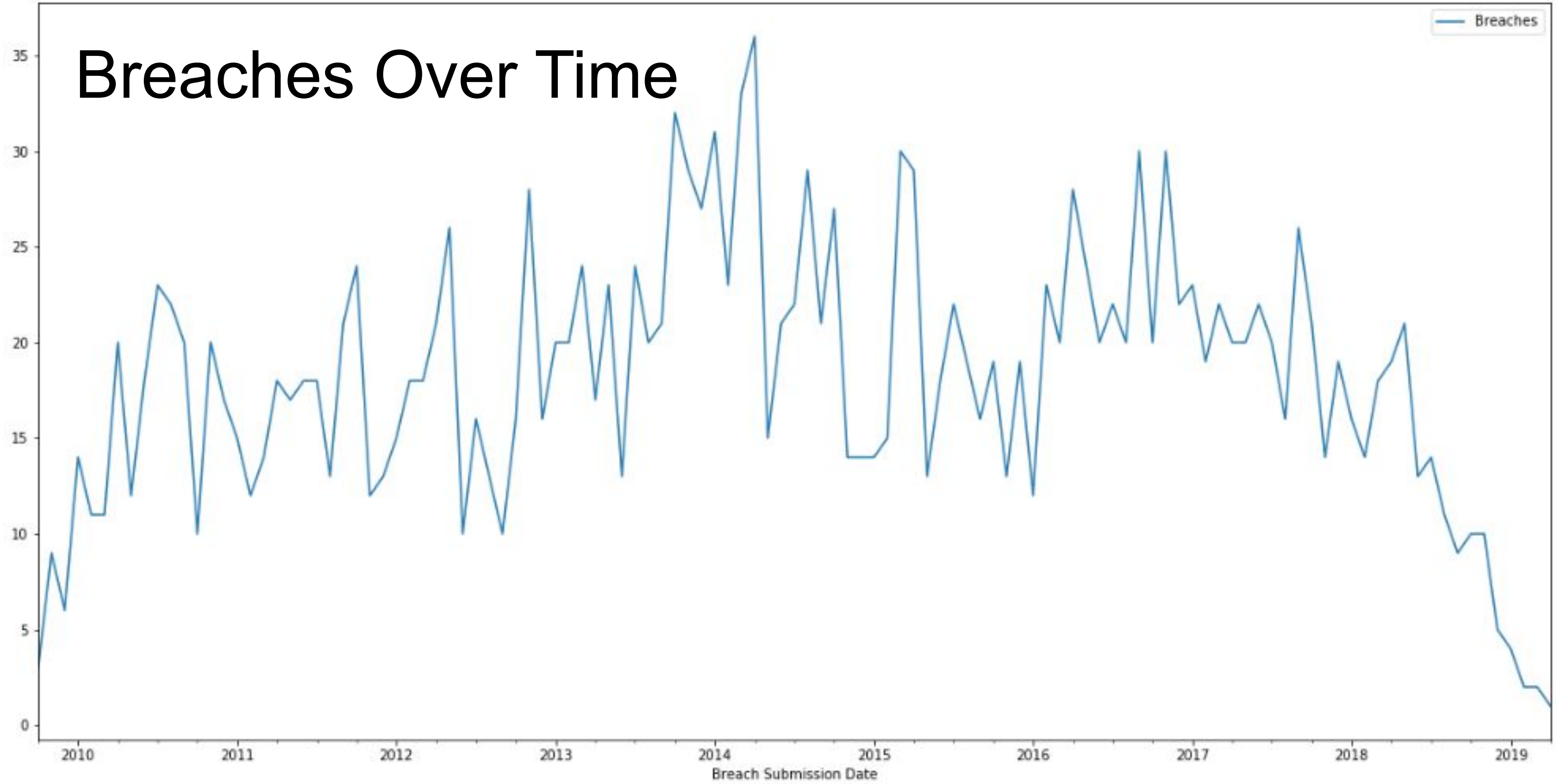


The top breach type is 'Theft' followed closely by "Unauthorized Access/Disclosure" and "Hacking/IT Incident" to make up the top three breach types.



The top data breaches are primarily electronic minus Paper/Films as the major data breach location type.

Breaches Over Time



Breaches increase from 2010 onwards, peaking around 2014 and then appear to follow a general decline.

Statistical Analysis

- Methods
  - T-test
  - Chi-squared test
- Variables
  - Breach type
  - Number Individuals affected
  - Covered Entity Type
  - Location of breach methods
- Findings
  - Relationship between breach type and the number of individuals affected
  - Relationship between breach type and location of breach method

Summary and Next Steps

- 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...

- Machine learning model
  - Random Forest
  - Naive Bayes
  - Decision Tree
  - Logistic Regression

In-Depth Analysis: Machine Learning Models

- **Data preprocessing:**
  - Dataset was balanced to have equal number of breach types (“*Hacking/IT Incident” vs. “other”*).
  - Further binary variables were created from categorical columns (“*Location of Breached Methods”, “Business Associate Present”*).
  - *“Breach Submission Date”* was split into “*year”, “month”, “date”, “weekend”, “week of year”*.
- **Models:**
  - Models used were Random Forest, Naive Bayes, Decision Tree, and Logistic Regression
- **Metrics & Tuning:**
  - Logistic Regression and Random Forest were the best performing models and were further tuned in order to optimize performance and determine important features.

Conclusion: Insights and Further Research

- ***Insights:***
  - Interesting features that modeling validated as being important in predicting breach type included *"Year", "Individuals Affected", " Week of Year".*
- ***Future work:***
  - Further research is needed in order to understand the importance of the features *"weeks"* and *"year"* in predicting type of breach.
  - More data is needed to see if there is a relationship between size of company and breach type.
  - 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