Machine Learning Foundations

Name: Georgina Woo
Date: Dec 9, 2023

This report summarizes my coursework for Break Through Tech's AI program at Cornell Tech.

1 Curriculum

Over the 9-week course, I attended weekly lab meetings consisting of a half-hour lecture, followed by two and a half hours of programming work. Each week, the deliverables included pre-class readings, a written assignment, a quiz, and a homework assignment based on the previous week's topic. The course introduced machine learning, with introductory lessons on data management and analysis with Python, followed by machine learning concepts such as K-Nearest Neighbors, decision trees, regression, and model building and analysis. The earlier coursework introduced us to common Python libraries used in data management, visualization, and machine learning such as pandas, matplotlib, seaborn, scikitlearn, and Pytorch, while most of the later coursework involved building and training machine learning models with given datasets to solve specific problems.

2 Final Project

2.1 Problem Statement

As a data scientist at X company, you work with a top international healthcare company. Leverage the accumulated data of past data breaches and develop a Machine Learning algorithm that predicts the method of the next data breach.

The dataset `df_1.csv` is a mostly numeric, tabular dataset that contains over 351 company data breach data totaling 30,000 records. There are 7 columns: Entity, year, records, organization type, method, sources, and index.

2.2 Approach

1) Data Understanding and Preprocessing
2) Exploratory data analysis
3) Feature Selection
4) Model Building
5) Model Evaluation

2.3 Dataset:

The csv was read as a pandas dataframe with the following columns:
`Entity:` Name of the company that was breached. This could be useful in future iterations of the project to identify if there are any patterns in the names of the companies that are breached using a Natural Language Processing approach.
`Year:` Year in which the breach occurred. Some entries are non-numeric.
`Records:` Number of records that were breached. Some entries are non-numeric.
`Organization type:` Type of organization that was breached. This is a categorical variable. Some entries are similar but have different names.
`Method:` Method of breach. This is a categorical variable. Some entries are similar but have different names.
`Sources:` Hyperlinked sources to the articles that were used to collect the data.
`Index:` Numerical index of the entry

2.4 Data Understanding and Preprocessing:

2.4.1 Organization Type

The organization types were modified to group similar types together. For example, `telcoms`, `telecom`, and `telecomms` were converted to the parent group `telecommunications`.

2.4.2 Records

Ill-formatted entries in the Records column were replaced with the mean of the records with the same organization type.

2.4.3 Years

The Years column had some ill-formatted entries. For example, some entries marked the years as 2004-05, 2004-06, etc. In these cases, the rows were duplicated for each year in the range, and the number of records was split evenly between the years.

```
     Entity   Year              Records     Organization type   Method

94   EasyJet  2019-2020         13394400    transport           hacked
96   Earl...  2018-2019         2000000     restaurant          hacked
144  Hilto... 2014 and 2015     363000      hotel               hacked

Null values:  False
Non numeric values:  True
```

2.4.4 Methods

The Methods column initially had 23 unique values, although many could be regrouped into the same class. Having the individual methods be separate created issues with training the model – especially on a relatively small dataset with 351 entries, there would not have been enough data to train a model to accurately predict one of 23 methods. For example, `unsecured_s3_bucket` could be classified under `poor_security`, and `stolen_laptop` and `lost/stolen laptop` could be classified under `stolen_media`. With 23 unique methods, the "hacked" entries lay as a statistical outlier (Fig 1), which resulted in accuracy issues during training and testing. However, dropping the outlier would result in losing half of the dataset.

I opted for reclassifying the methods into one of 5 parent methods as the subsets of the 5 classes that result from the re-classification can be addressed with the same recommendations (Fig 2). For example, if a company is vulnerable to hacking from an active attacker, we can suggest they hire a white hat firm to probe their systems, and can then make recommendations on vulnerabilities within their system. So, they can prepare their systems by seeking recommendations from network security specialists. Meanwhile, if a company is likely to lose its data through poor security or lost media, this can be managed with mandates to encrypt all their data, especially on removable disks like laptops and USB sticks, as well as more internal training for best practices in IT habits.

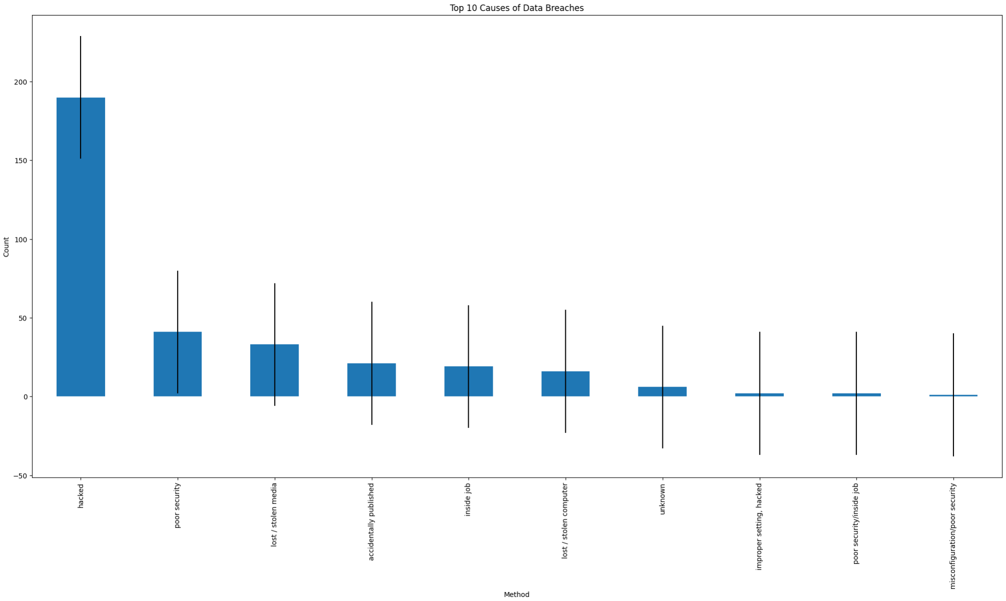

Figure 1: Before Re-classification

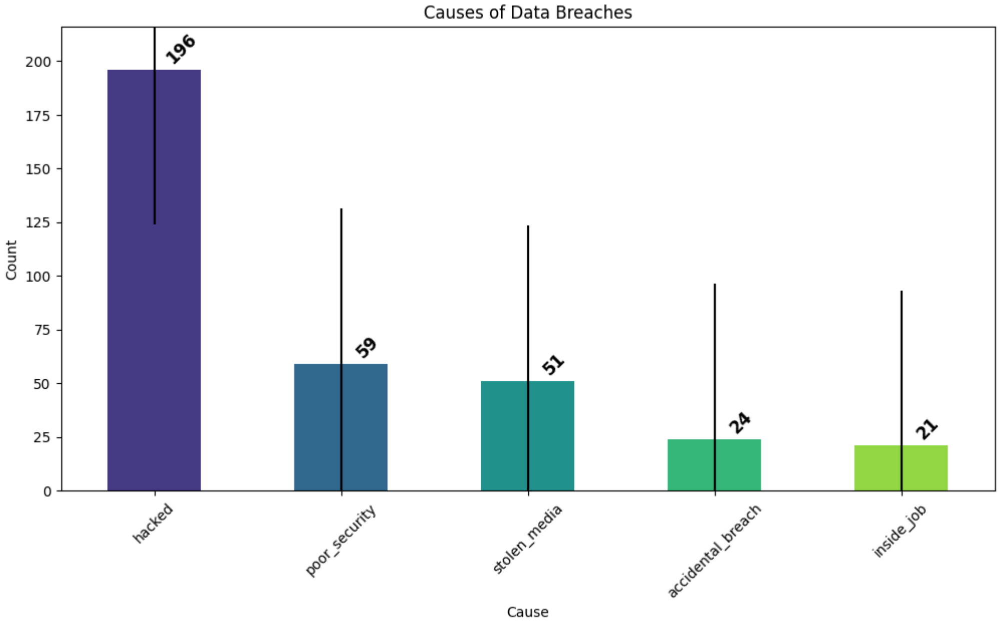

Figure 2: After Re-classification

2.5 Exploratory Data Analysis

The dataset might not be fully representative of real-world data breaches – for example, looking at the healthcare industry alone, there was no healthcare data lost in half of the years. Some research shows that this is not the case, although the spike in data lost in 2014-2015 is accurate, as well as the steady increase thereafter.

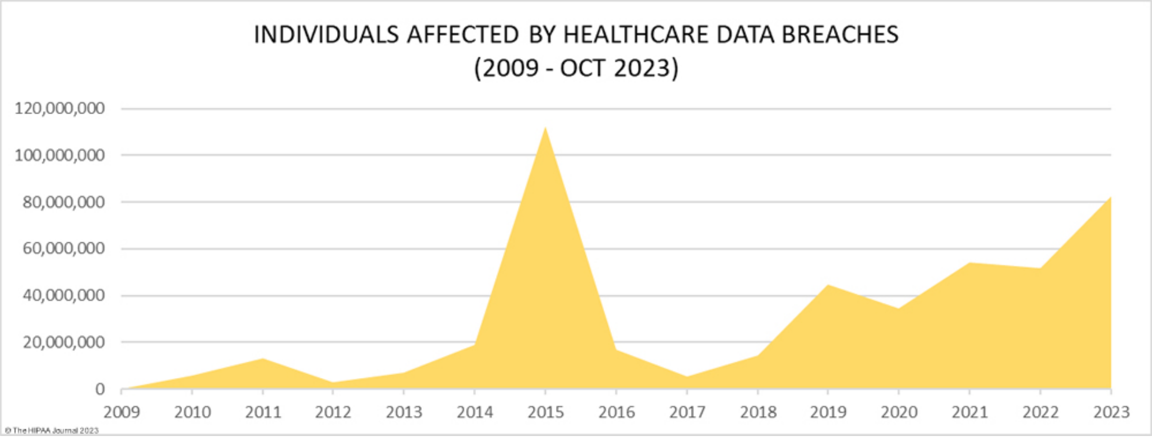

Figure 3: Source: https://www.hipaajournal.com/healthcare-data-breach-statistics/

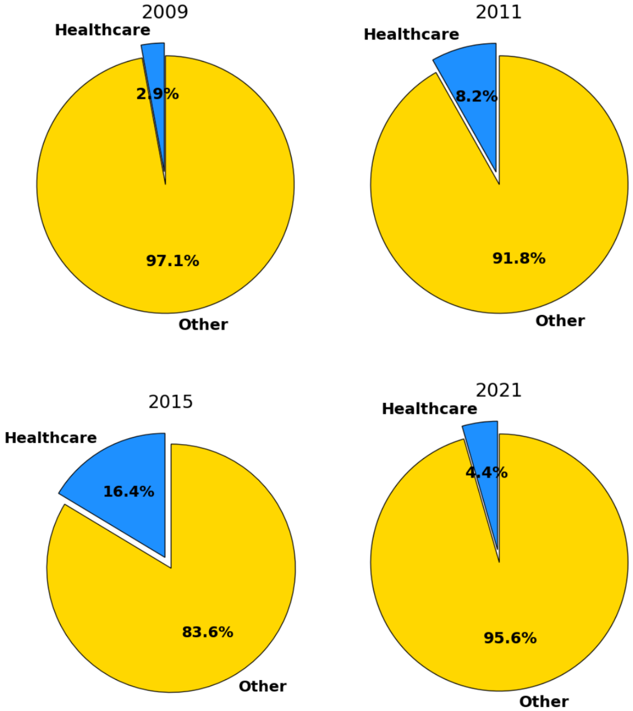

Figure 4: Given data on healthcare data breaches

This could be because the dataset is not fully representative of all data breaches that have occurred in the past 15 years. It is also worth noting that the range of entries in the dataset is between 2004 and 2019, so more recent data breaches are not included.

In both the real-world data and our given dataset, The loss/theft of healthcare records and electronically protected health information dominated the breach reports between 2009 and 2015. The move to digital record keeping, more accurate tracking of electronic devices, and more widespread adoption of data encryption have been key in reducing these data breaches.

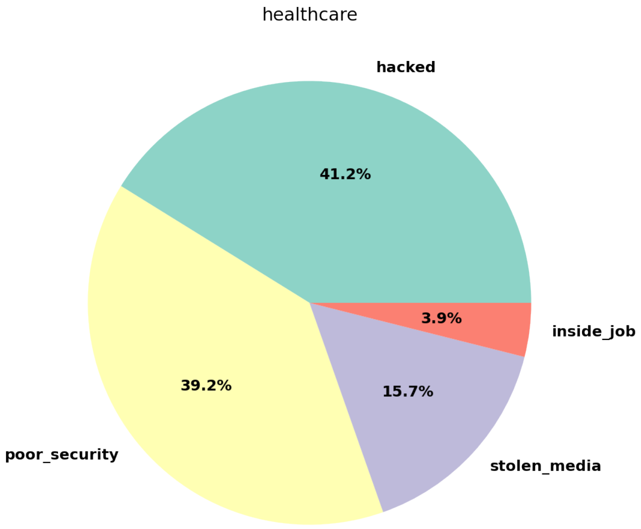

Figure 5: Distribution of healthcare data breaches

2.6 Feature Selection

The Sources and Index columns were dropped, and I used the Entity, Organization Type, Year, and Records as features.

2.7 Model Building

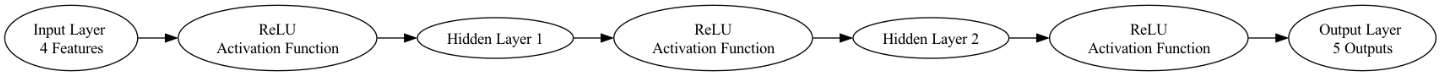

Figure 6: Neural Network

The values from the dataframe are encoded with integer values with a label encoder. After standardizing the features and using Pytorch to convert both the features and label into Pytorch tensors. I used Pytorch's nn.Sequential module, choosing to use two hidden layers on top of the input and output layer, with ReLU activation functions. The hidden layers and the activation function transform the input data and introduce non-linearity to the network to capture different features or representations of the input data, meaning it can model complex relationships between the data, and learn and represent intricate patterns.

For model training, I've trained the model on the training sets iteratively and used cross-validation. For each fold the model is split into a training and validation set, where in training, for each epoch the model sees the training example once and updates its weight parameters with backpropagation based on the computed loss. The validation set is then used to assess the model's performance. I implemented an early stopping feature where if the validation loss doesn't improve after 2 iterations, that training loop is stopped early, which allowed me to find the epoch value of 75 for the best performing model.

2.8 Model Evaluation

This was a classification model to predict the method of the data breach. The model performed with an 81.82% accuracy rate, along with a validation loss of 0.29, which measures how the model performs on data it has not seen during its training.