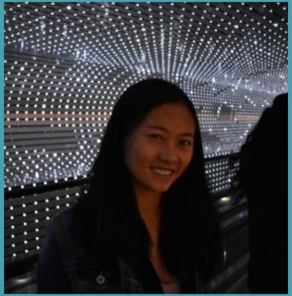
InnSure

Northeastern
D'Amore-McKim School of Business
DATA Initiative

Predicting Data Breaches Using Human Factors

Spring 2022

ABOUT US


Julie Corfman
Project Manager

Will Gandre
Relationship Manager

Nhat Pham
Chief Data Scientist

Yuhan Qiu
Literature Expert

TABLE OF CONTENTS

1 **PROJECT GOALS**

Data Gathering & Breach Prediction Model

2 **PRIOR FINDINGS**

Human Factors May Signal Cyber Resilience

3 **TOOLS & RESOURCES**

Breach Databases, People Databases & Social Networks

4 **METHODOLOGY**

Web Scrapers & Predictive Models

5 **RESULTS**

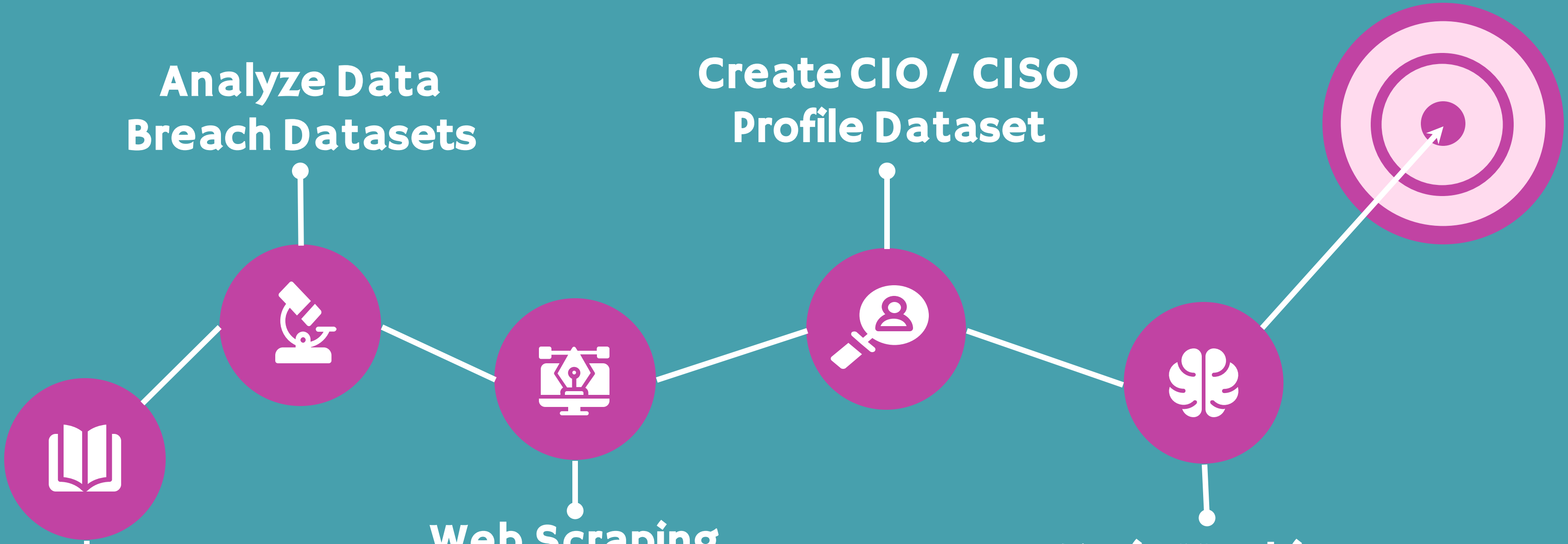
Random Forest and AdaBoost Yield Encouraging Results

6 **INSIGHTS & WRAP-UP**

Experience Stands Out As Influential Marker

PROJECT GOALS

PROJECT GOALS



Literature Review

Analyze Data Breach Datasets

Web Scraping Capability

Create CIO / CISO Profile Dataset

Train Machine Learning Model

Breach Risk Index

PRIOR FINDINGS

CONTEXT

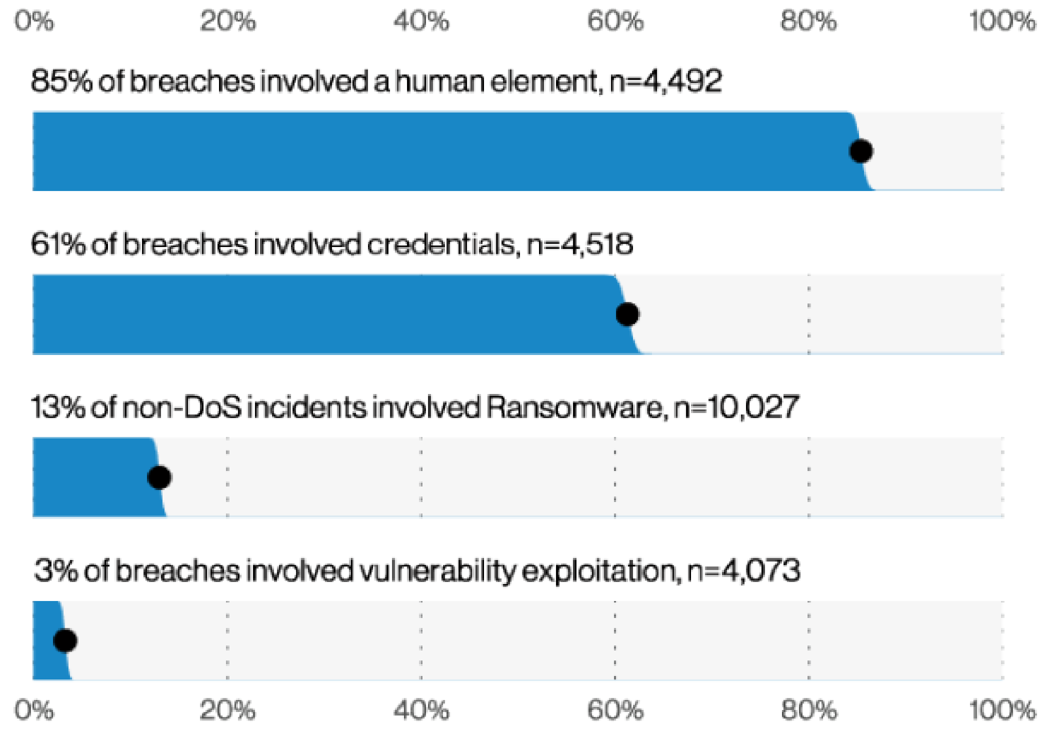
Over 1,000

Annual publicized data breaches
(Statista, 2020)

$4.24 Million

Average total cost of a data breach
(IBM, 2021)

CONTEXT



(Verizon, 2021)

PREVIOUS GROUP

TWO HEALTH SYSTEMS

Same industry and of comparable size

MULTIPLE PROFILES

Members of the board, C-suite, and IT upper management.

CLASSIFICATION

Logistic regression for binary prediction of data breaches

FINDINGS

Education and experience appear to be predictive factors; larger dataset needed to verify

TOOLS & RESOURCES

TOOLS & RESOURCES

VERIS & PRC

Community driven & non-profit data breach datasets

Phantombuster

Catalog of data extraction tools that work on popular web and social media sites

LinkedIn

Professional networking site that includes members' self-reported CV's

People Data Labs

B2B vendor of "People Data"

METHODOLOGY

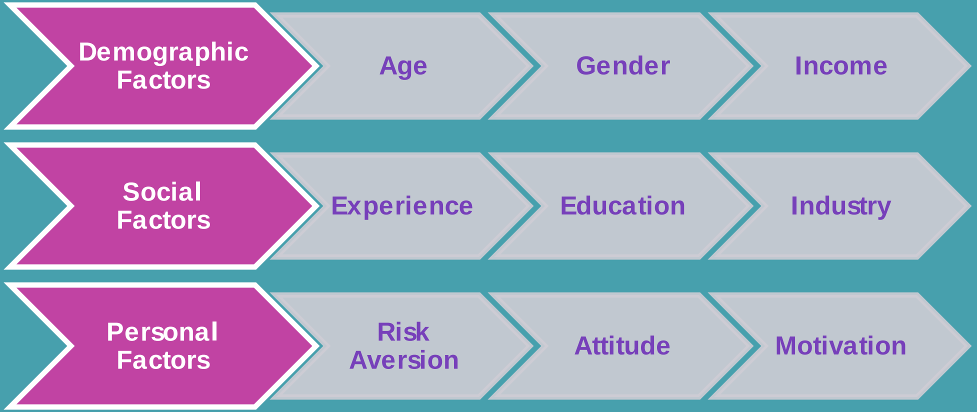
ACADEMIC LITERATURE

Human Factors

Machine Learning Models

LITERATURE REVIEW

HUMAN FACTORS RELATED TO CYBER SECURITY



Demographic Factors	Age	Gender	Income
Social Factors	Experience	Education	Industry
Personal Factors	Risk Aversion	Attitude	Motivation

- Employee, contractors and vendors are responsible for the majority of data breaches (Verizon, 2021; Bailey, 2018)
- Sociodemographic factors are correlated with cyber security knowledge and behavior (Prabhu, 2021; Zaman, 2020)
- CIO and executive characteristics influence the likelihood of a data breach (Haislip, 2021; Smith 2021)

LITERATURE REVIEW

FORECASTING CYBER INCIDENTS USING MACHINE LEARNING MODELS

- Supervised machine learning models have been used to predict data breaches and cyber hacking:
  - Tree-based models: Random Forest
  - Logistic Regression
  - Support Vector Machine
- Used public data breach datasets which included the Vocabulary for Event Recording and Incident Sharing (VERIS) and Privacy Rights Clearinghouse (PRC)
- Technical factors, reputation on Twitter and organizational properties were used as inputs (see graph)—human factors were not

LITERATURE REVIEW

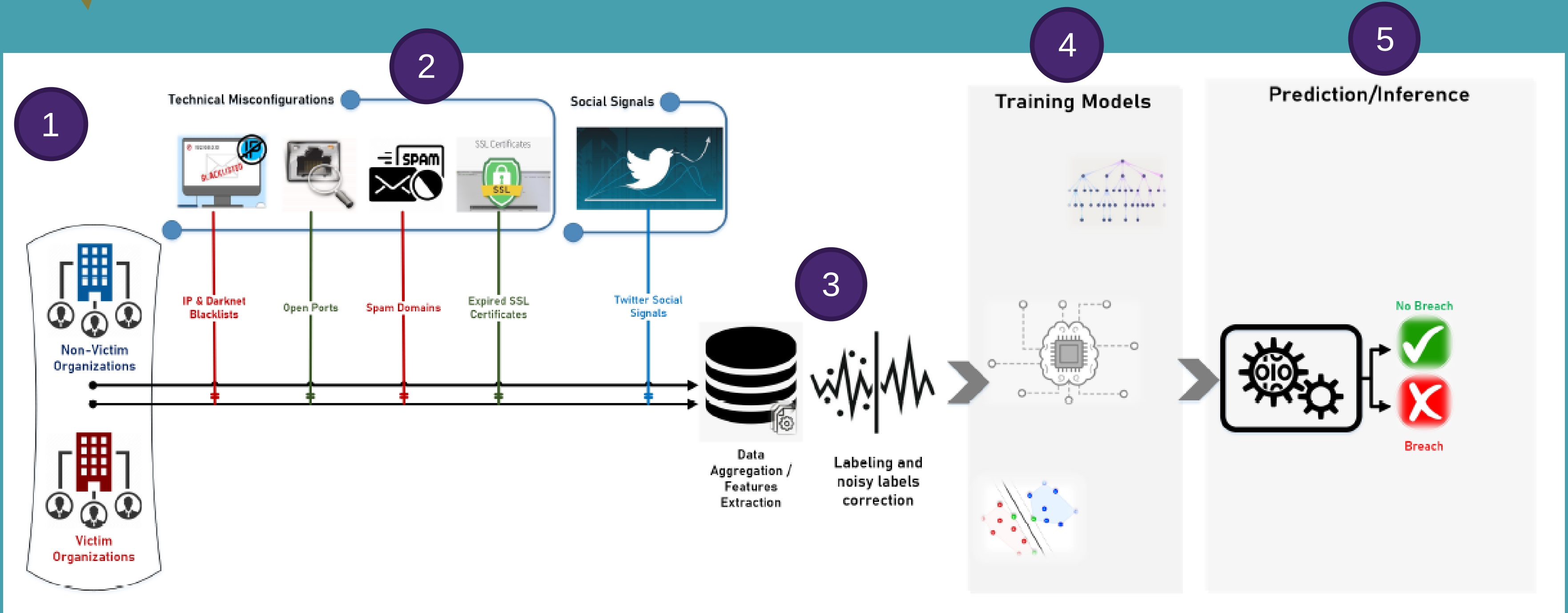


Fig. 1. STRisk pipeline to combine technical misconfigurations and Twitter social signals for both victim and non-victim organizations, correct noisy labels and build the predictive models to discriminate risky organizations from non-risky ones

(Benbrahim, 2021)

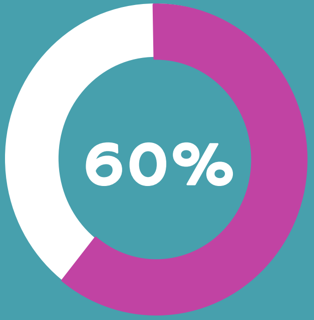
Data Breach Datasets

Vocabulary for Event Recording and Incident Sharing (VERIS)

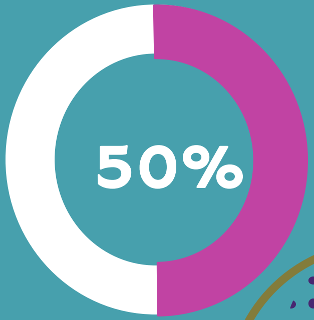
- Includes self-reported incidents and data breaches, as well as publicized breaches (N = 10,324)
  - ➢ Incident: No proof of data disclosure to unauthorized party
  - ➢ Breach: Data viewed, accessed or downloaded by an unauthorized party
- Dataset weighted towards health care and public sector due to public reporting laws
- Breach records used to select breached organizations and filter those selected as non-breach organizations

Privacy Rights Clearinghouse (PRC)

- Exclusively data breaches that are of public record (N = 9,015)
- Records used to filter non-breach organizations; lacked appropriate meta-data to select appropriate breached organizations

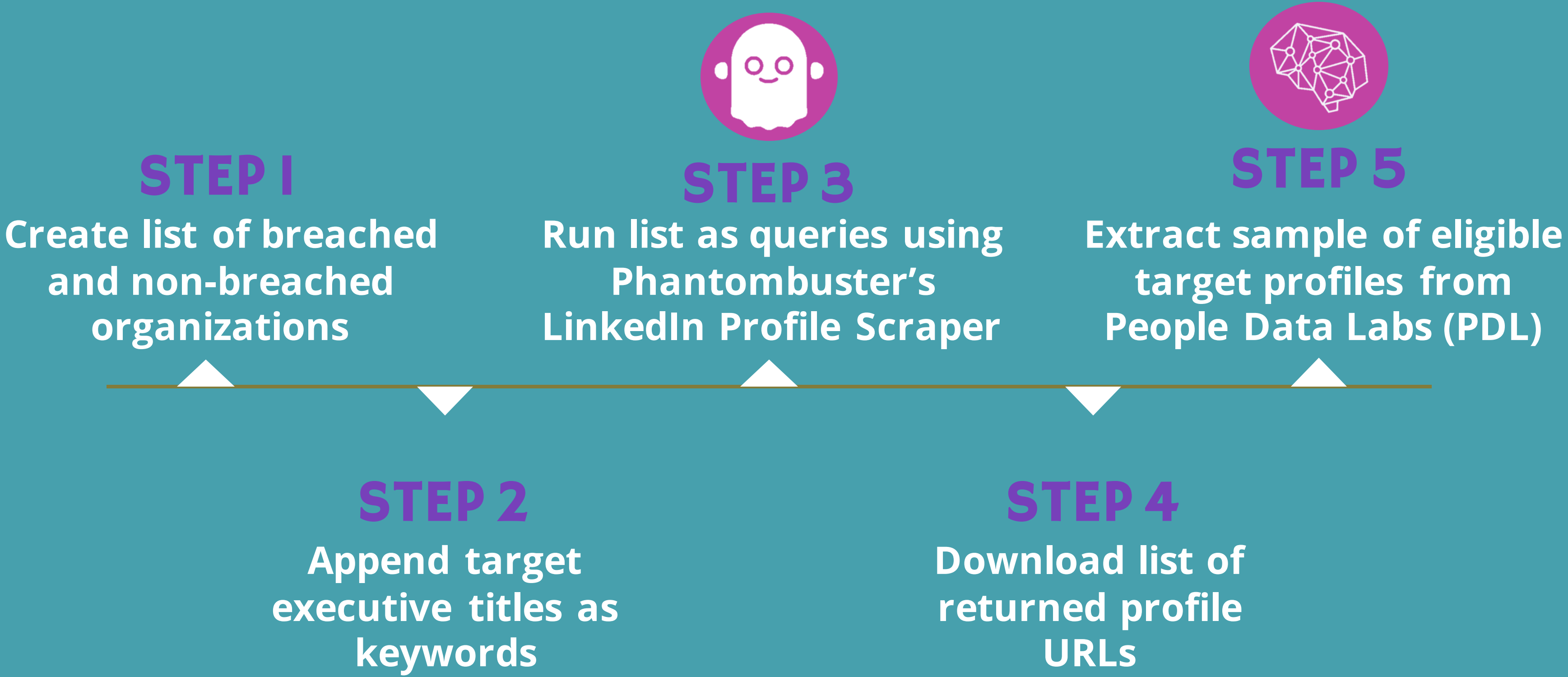


VERIS records categorized as a breach



VERIS breaches from the health care or public sector

WEB SCRAPING



STEP 1

Create list of breached and non-breached organizations

STEP 2

Append target executive titles as keywords

STEP 3

Run list as queries using Phantombuster's LinkedIn Profile Scraper

STEP 4

Download list of returned profile URLs

STEP 5

Extract sample of eligible target profiles from People Data Labs (PDL)

EXECUTIVE PROFILE DATASET

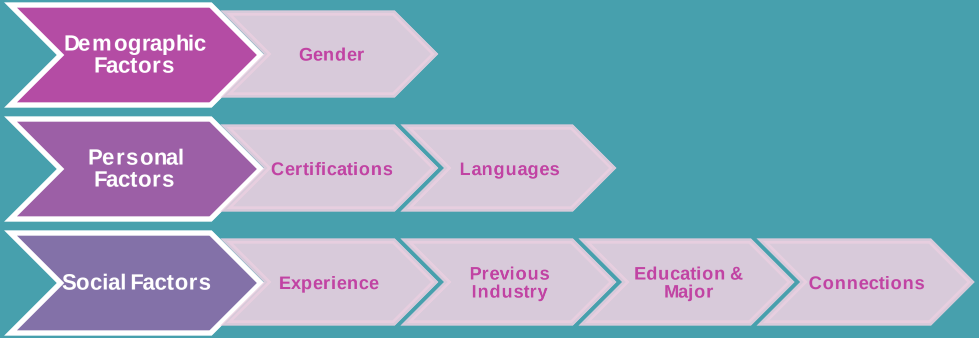
(raw data)

671 Observations, **277** Columns
345 Breach vs **326** Non-Breach

Basic Information:

Industry (77)
Company Size (1-10000+)
Job Title

Human Factor Related Information



Demographic Factors: Gender

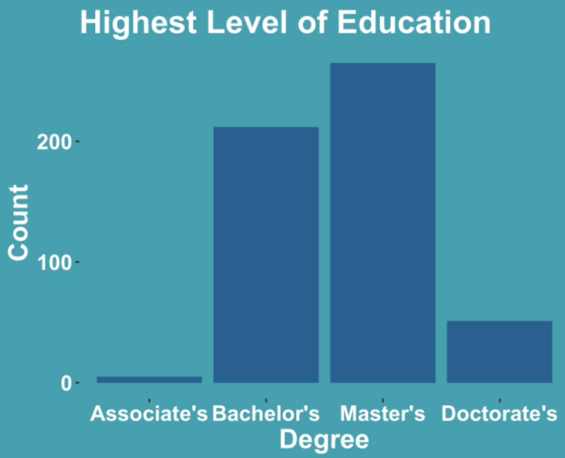
Personal Factors: Certifications, Languages

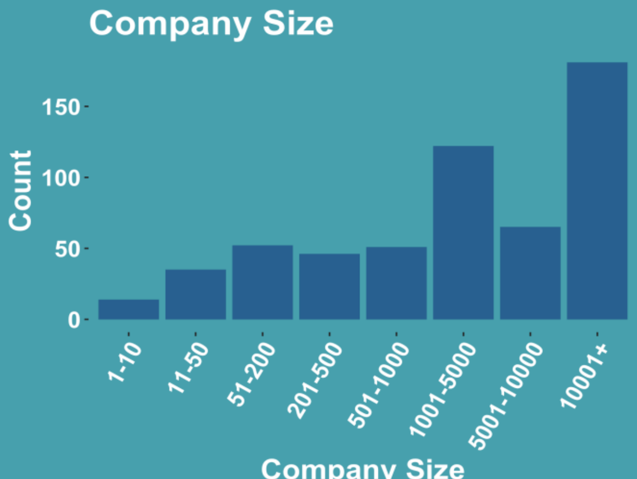
Social Factors: Experience, Previous Industry, Education & Major, Connections

EXECUTIVE PROFILE DATASET (CONT.)

Industry	Count
Hospital & Health Care	153
Information Technology and Services	99
Financial Services	37
Government Administration	35
Insurance	27

Title	Count
Chief Information Officer	173
Chief Technology Officer	69
Chief Information and Security Officer	61
Vice President and Chief Information Officer	17
Information Security Officer	16





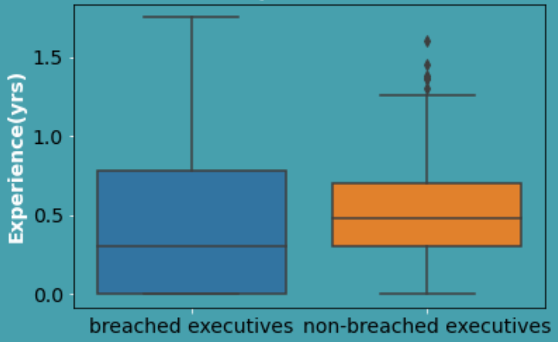
PREDICTIVE MODEL

A. Data Preprocessing

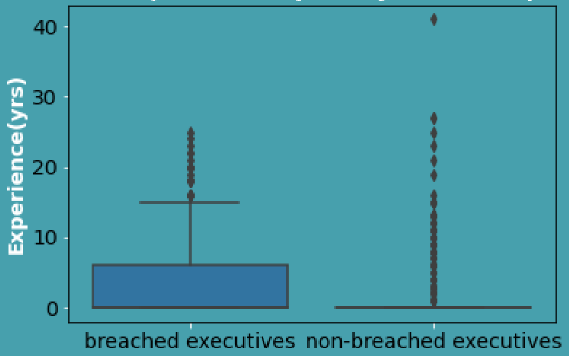
- From scraped data, insights from literature review and discussions with client, we aggregated **96 human-related features** for modeling:
  - **Education –** highest level of education, major, degrees, having MBA, JD, MD degrees, number of years since last education degree
  - **Working Experience –** years of working experience, years of experience until breach, number of years in nonprofit/ private/ public/ governmental sector
  - **Skills –** 59 leadership/business and technical skills
  - **Information Security Certifications –** 8 popular certificates
  - **Other –** gender, number of LinkedIn connections, number of languages, number of social media accounts available

PREDICTIVE MODEL

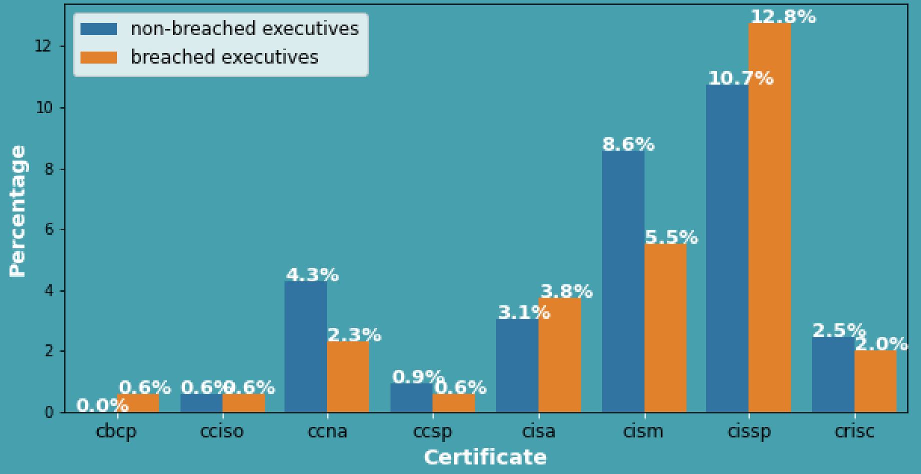
B. Exploratory Data Analysis



Years of experience till breach



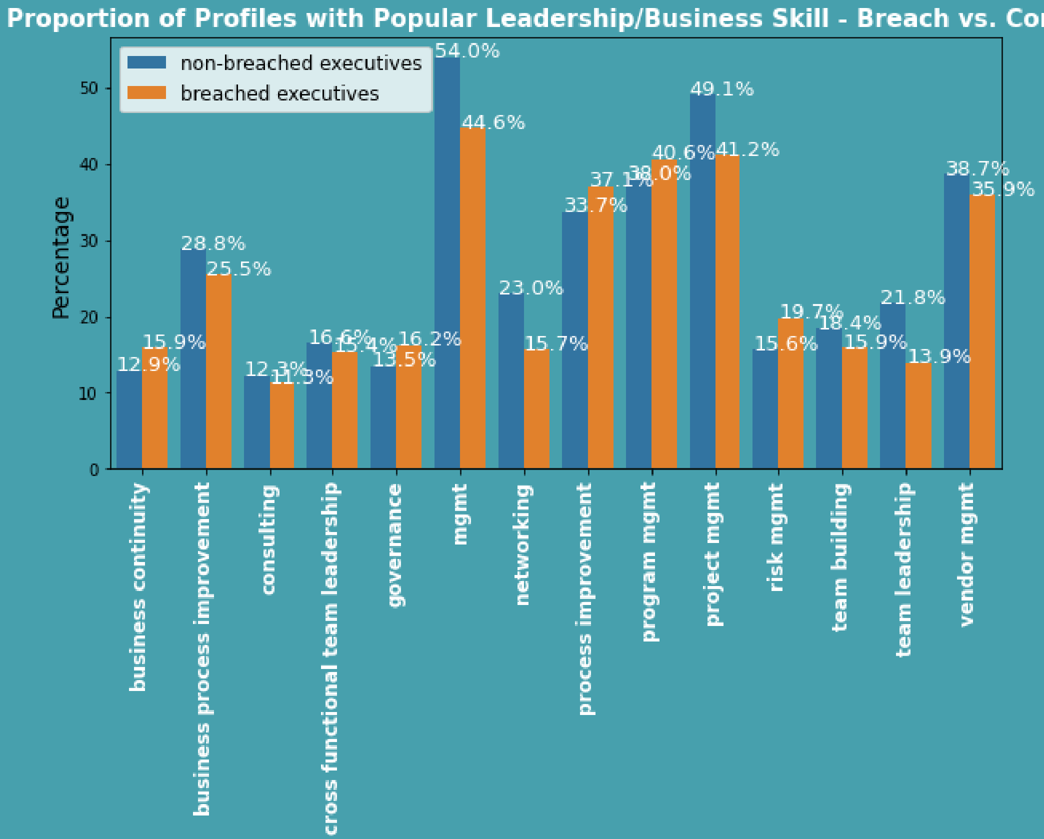
Years of experience in publicly listed companies



Proportion of Profiles with InfoSec Certificates - Breach vs. Control

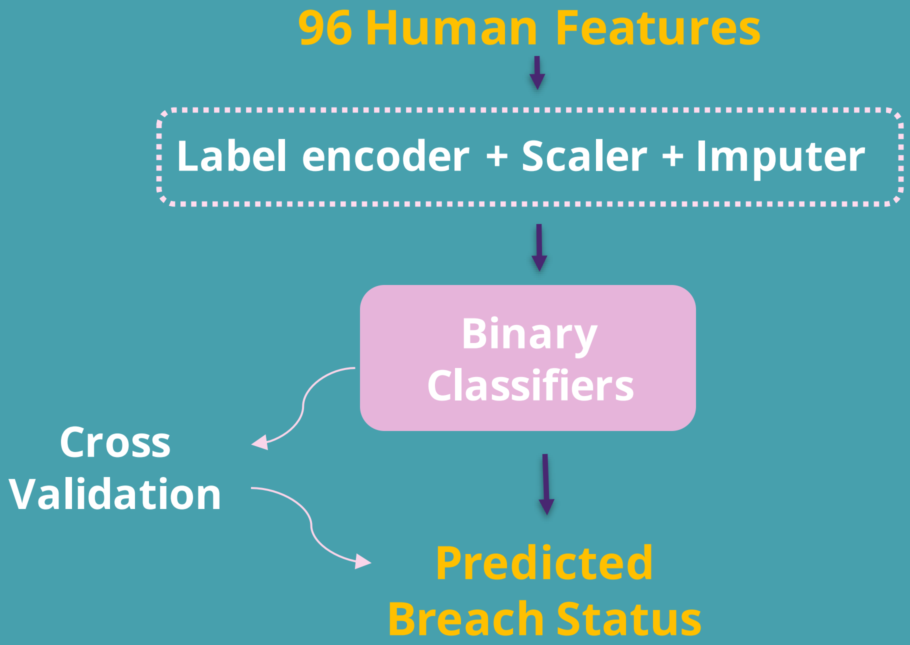
PREDICTIVE MODEL

B. Exploratory Data Analysis



PREDICTIVE MODEL

C. Classification Models



96 Human Features

Label encoder + Scaler + Imputer

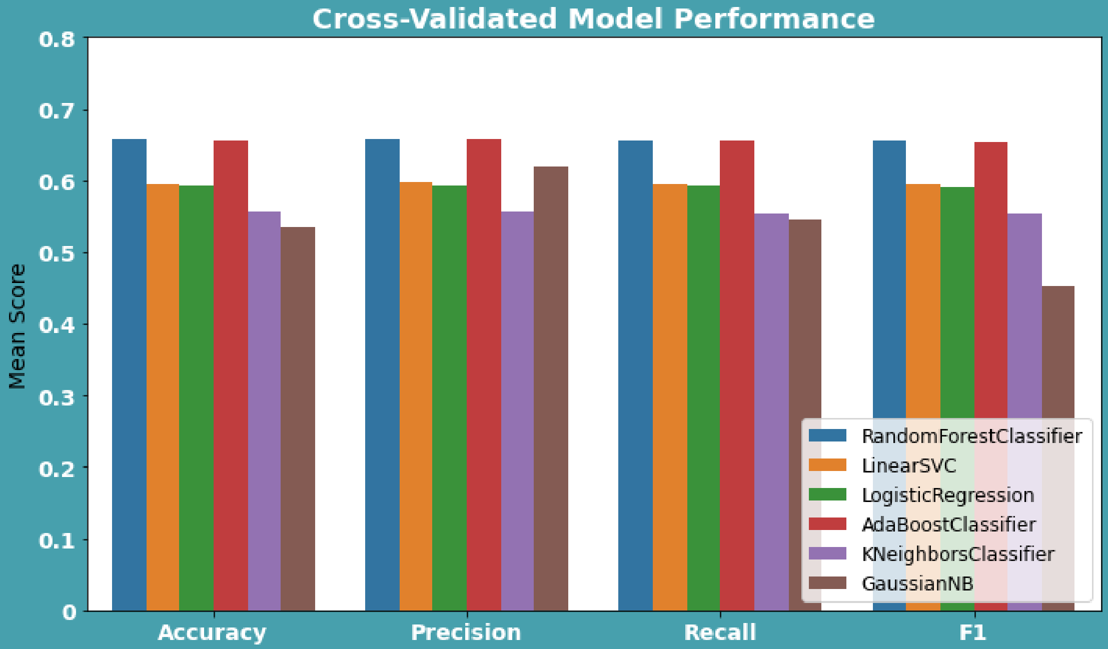
Binary Classifiers

Cross Validation

Predicted Breach Status

PREDICTIVE MODEL

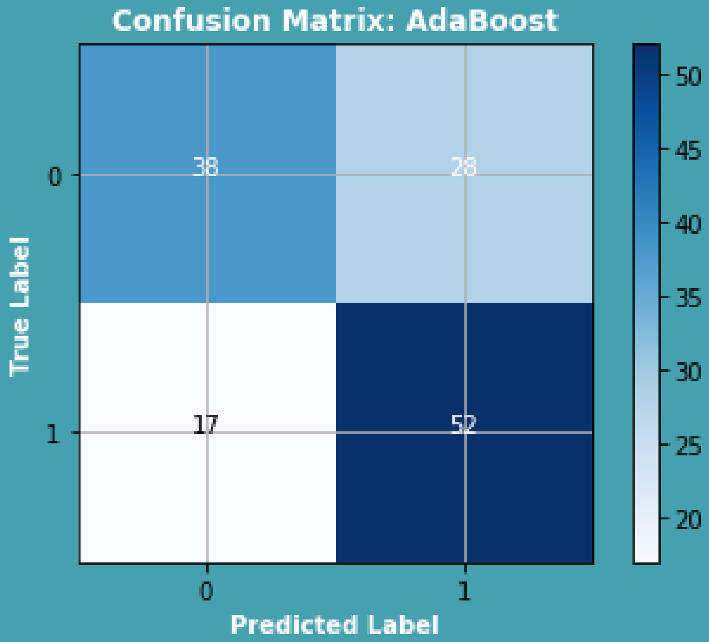
- Cross validation results show no sign of overfitting

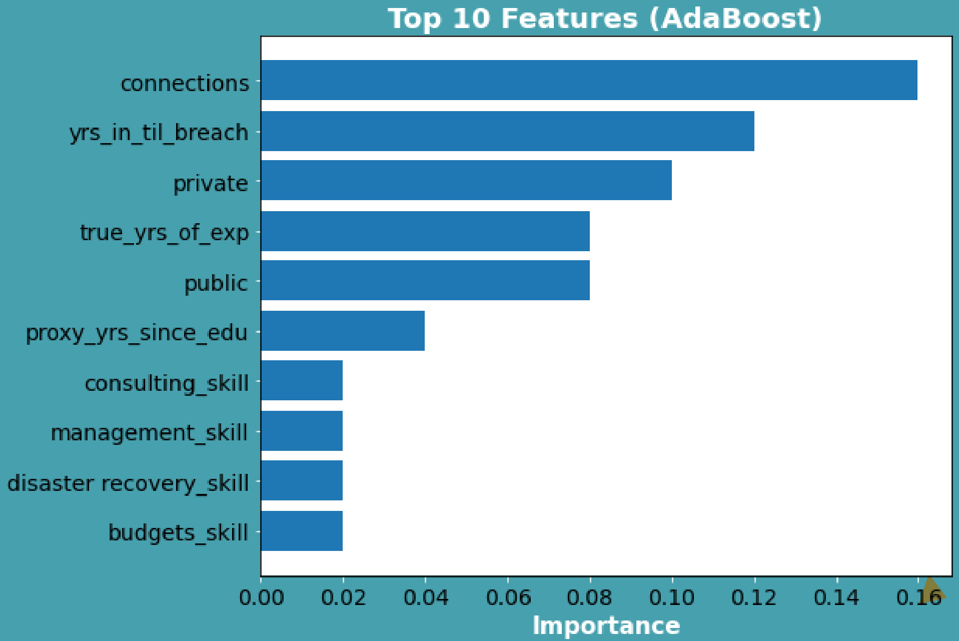


PREDICTIVE MODEL

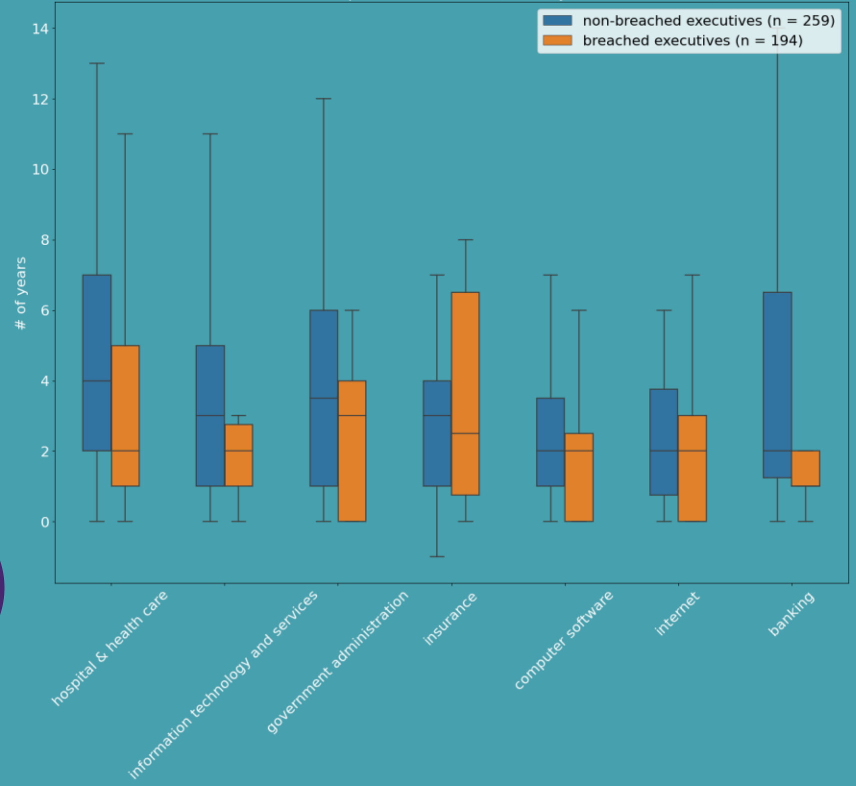
C. Results

- AdaBoost and Random Forest classifiers performed the best with ~67% accuracy

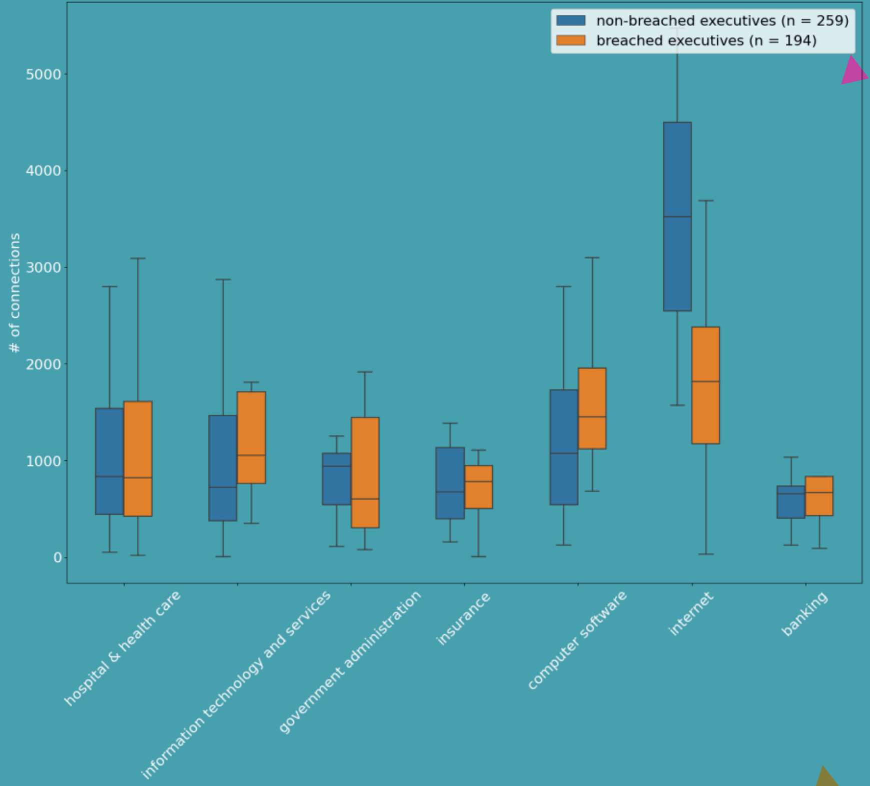




YEARS IN POSITION & CONNECTIONS



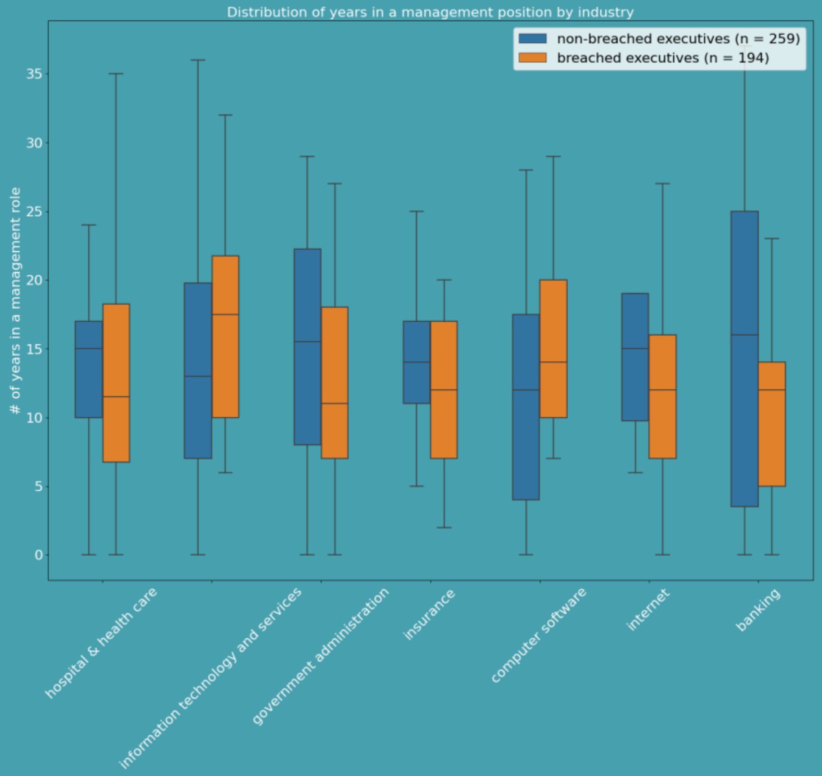
Distribution of Years in position until breach (or # of years in CIO level role)

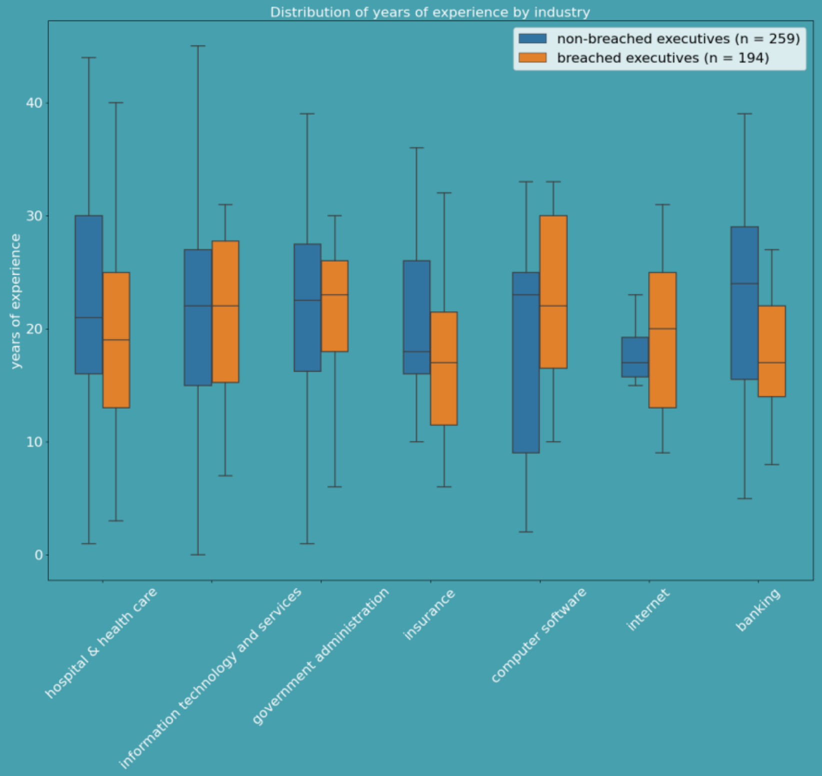


Distribution of # of LinkedIn Connections

Note: For non breached executives, the # of years until breach is represented as # of years in the CIO level role for control company

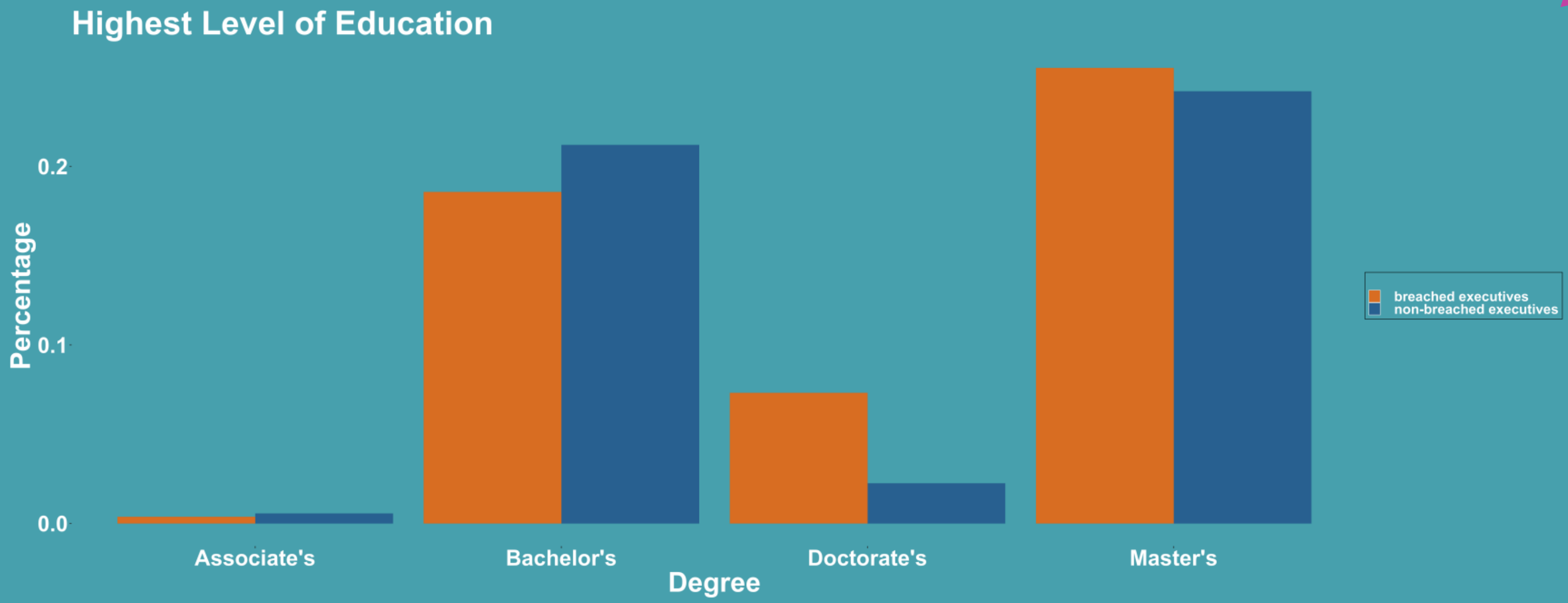
YEARS OF EXPERIENCE



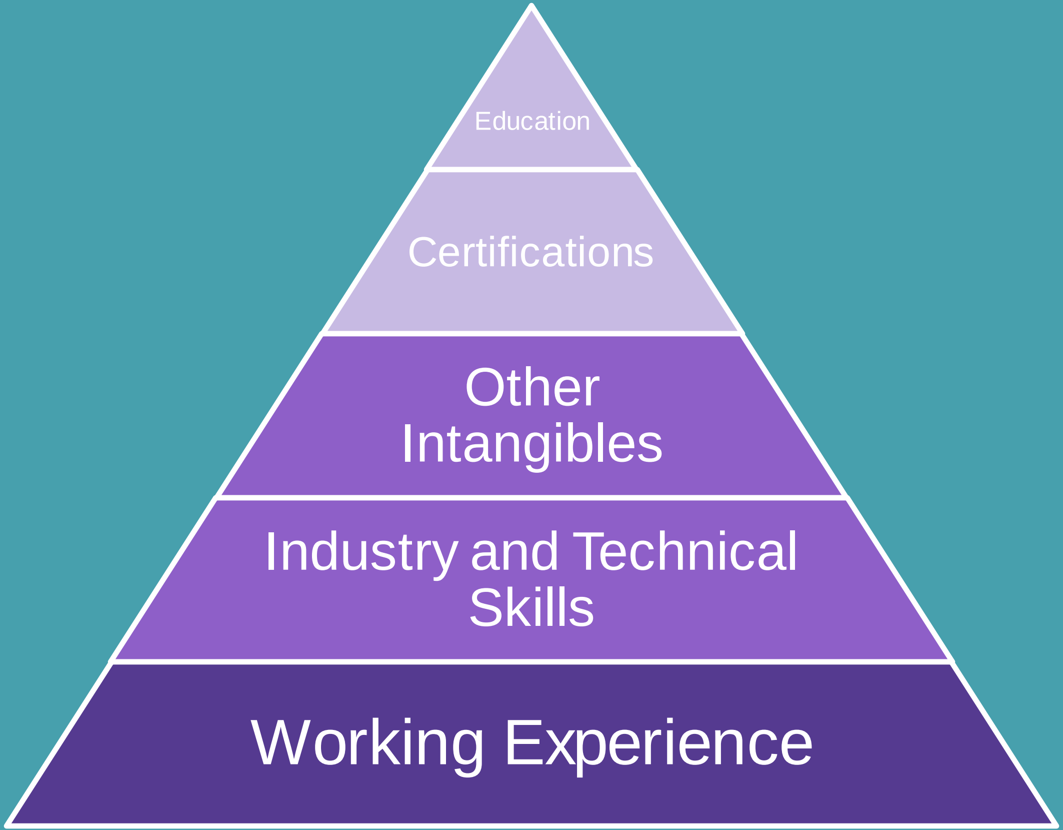


Note: Left: years in a director, C-level, or manager role; Right: total years of experience (including possible precollege experience); outliers removed

EDUCATION


Highest Level of Education
Percentage
0.2
0.1
0.0
Associate's
Bachelor's
Doctorate's
Master's
Degree
🟧 breached executives
🟦 non-breached executives

Predictive Power Pyramid



Education

Certifications

Other Intangibles

Industry and Technical Skills

Working Experience

RECAP

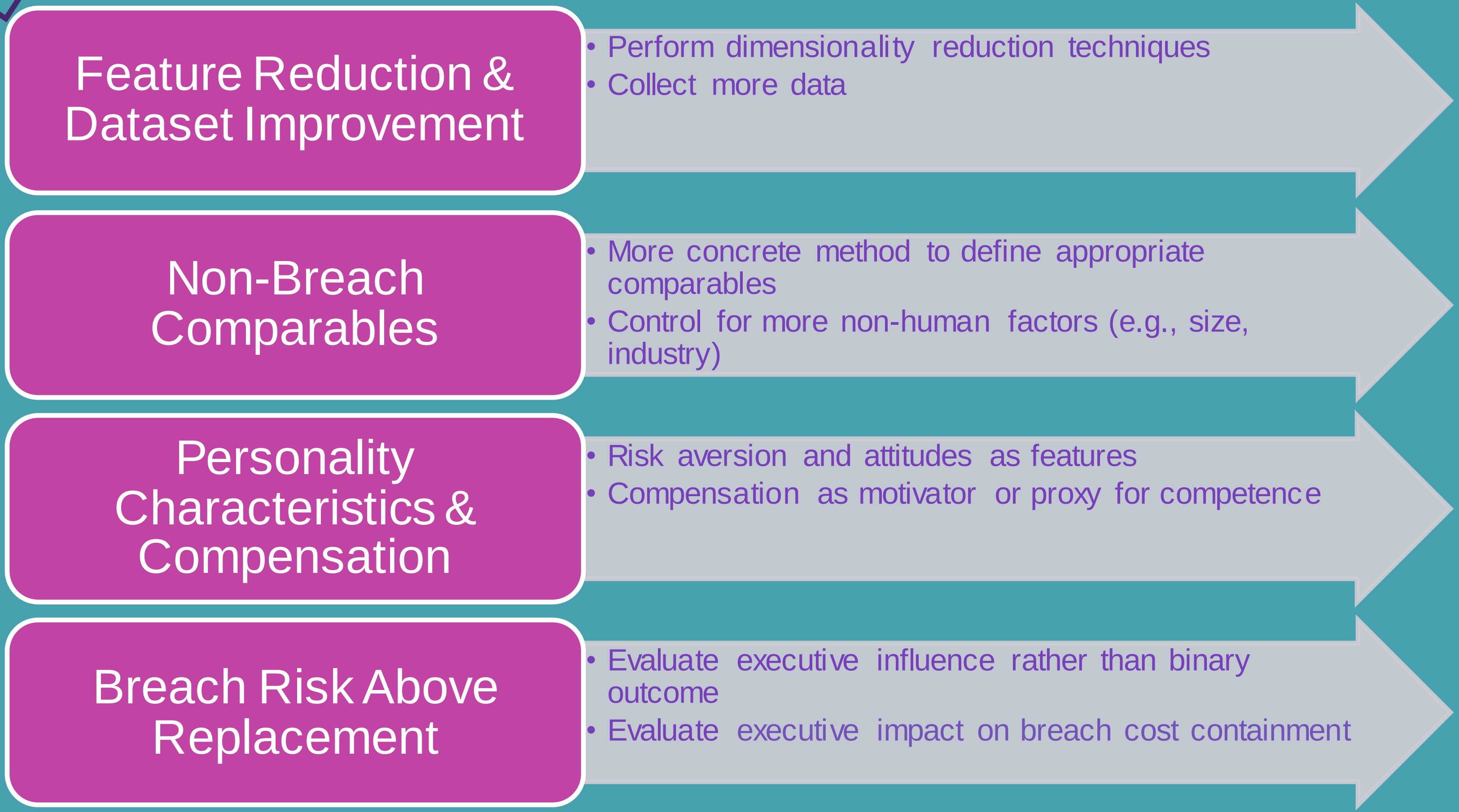
Our findings

- When the data breach occurred, **executives at breached companies were generally in their roles for fewer years** than executives at non-breached companies
- Executives from **non-breached companies typically have fewer connections per industry** (except for internet).
- Among non-breached executives, **the median number of years spent in a management position is higher**

What we did not find

- In our dataset, **certificates advertised on LinkedIn or education have little predictive value**

FUTURE WORK



Feature Reduction & Dataset Improvement

- Perform dimensionality reduction techniques
- Collect more data

Non-Breach Comparables

- More concrete method to define appropriate comparables
- Control for more non-human factors (e.g., size, industry)

Personality Characteristics & Compensation

- Risk aversion and attitudes as features
- Compensation as motivator or proxy for competence

Breach Risk Above Replacement

- Evaluate executive influence rather than binary outcome
- Evaluate executive impact on breach cost containment

Questions?

References

Benbrahim, H., Ghogho, M., Hammouchi, H., Mezzour, G., & Nejjari, N. (2021). STRisk: A Socio-Technical Approach to Assess Hacking Breaches Risk. *IEEE Transactions on Dependable and Secure Computing*. https://doi.org/10.1109/TDSC.2022.3149208

Bateman, R. M., Schweitzer, K. M., Xu, M., & Xu, S. (2018). Modeling and Predicting Cyber Hacking Breaches. *IEEE Transactions on Information Forensics and Security*, 13(11), 2856–2871. https://doi.org/10.1109/TIFS.2018.2834227

Haislip, J., Lim, J.-H., & Pinsker, R. (2021). The Impact of Executives' IT Expertise on Reported Data Security Breaches. *Information Systems Research*, 32(2), 318–334. https://doi.org/10.1287/isre.2020.0986

References

Prabhu, S., & Thompson, N. (2021). A primer on insider threats in cybersecurity. *Information Security Journal*, 30(1), 1-10. https://doi.org/10.1080/19393555.2021.1971802

Liu, Y., Naghizadeh, P., Sarabi, A., Zhang, J. (2015). Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents. *Proceedings of the 24th USENIX Security Symposium*. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/liu

IBM. (2021). Cost of a Data Breach Report. https://www.ibm.com/security/data-breach

References

Smith, T., Tadesse, A. F., & Vincent, N. E. (2021). The impact of CIO characteristics on data breaches. *International Journal of Accounting Information Systems*, 43, 100532. https://doi.org/10.1016/j.accinf.2021.100532

Verizon. (2021). DBIR 2021 Data Breach Investigations Report. https://www.verizon.com/business/resources/reports/dbir/

Zaman, S. (2020). The Effects of Human Factor Dynamics in Cyber Security in Kuwait. 3rd IET International Smart Cities Symposium, 3SCS-2020. https://ieeexplore.ieee.org/document/9545493