Predicting Cybersecurity Incidents Through Mandatory Disclosure Regulation

Aniket Kesari
JD/PhD Student, Jurisprudence & Social Policy
Yale University/The University of California, Berkeley

October 10, 2020

Abstract

Cybercrime is an increasingly common risk for organizations that collect and maintain vast troves of data. There is extensive literature that explores the causes of cybercrime, but relatively little work that aims to predict future incidents. In 2011, the United States Securities and Exchange Commission (SEC) provided guidelines for how publicly traded companies should convey these risks to potential investors. The SEC and other regulatory agencies are exploring how to leverage artificial intelligence, machine learning, and data science tools to improve their regulatory efforts. This paper explores the potential to use machine learning and natural language processing techniques to analyze firms' mandatory risk disclosure statements, and predict which firms are at the greatest risk of suffering cybersecurity incidents. More broadly, this study highlights the potential for using legally mandated disclosures to bolster regulatory efforts, particularly in the context of prediction policy problems.

1 Introduction

"Sunlight is said to be the best of disinfectants, electric light the most efficient policeman.", Louis Brandeis in *Other People's Money and How the Bankers Use It* (1914)

In 1914, Louis Brandeis wrote this powerful statement in response to the emergence of consolidated banks and trusts (Brandeis, 1914). He was concerned about the power these institutions would have in American democracy, and prescribed several solutions. Among these was the notion of "sunlight as disinfectant" - that transparency and openness were effective means to regulate these large enterprises that could perpetuate a range of social ills. At the time, Brandeis called for the creation of a government agency that could force transparency and investigate wrongdoings. These ideas were the foundation that formed what became the Federal Trade Commission.

Today, large corporations deal not only with other people's money, but also their data. Over the last several decades, the U.S. adopted several data protection laws that regulate particular economic sectors that deal with especially sensitive data. Mandatory disclosures are a popular tool

for encouraging good corporate behavior. "Sunlight as disinfectant" is the main theory underlying mandatory disclosure laws, and the notion is that consumers and regulators can punish corporations that engage in bad behavior. So long as there is adequate information, the public and government agencies are well positioned to prevent the types of social ills that stem from consolidated corporate power.

Cybersecurity incidents that result in the loss of consumer data, especially losses attributable to external breaches, pose a serious threat to consumer privacy. Such incidents are becoming more severe, as evidenced by recent news headlines surrounding Cambridge Analytica (Lapowsky, 2018) and the Equifax breach (Cowley, 2017). These most recent events each implicated at least 80 million records, and their unprecedented scale has prompted policymakers at both the federal and state levels to consider and pass legislation to prevent future events. Academics have seriously engaged in the theoretical, technical, economic, and policy dimensions underlying privacy and cybercrime for decades, but there remain a number of open empirical questions.

The cybercrime literature is largely concerned with detecting and deterring cybercrime. However, relatively little attention has been paid to predicting incidents of cybercrime. Compared to traditional crime, cybercrime's spatial dimensions are difficult to conceptualize, and cybercrimes are somewhat rare events. Despite this difficulty, successfully predicting cybercrime could potentially yield enormous benefits. Compensating victims of cybercrime for their losses after a breach is difficult because it is hard to measure the damage (Mayer, 2016). Finding and punishing the perpetrators of cybercrime may be nearly impossible, especially if they live in a non-U.S. jurisdiction. However, deterring cybercrime by making it economically impractical may be more effective. Much of cybercrime is financially motivated, and choking off financial incentives is a powerful way to deter it.

In this piece, I propose predicting incidents of cybercrime primarily by looking at the potential risk factors a company may exhibit. Cybercriminals may exploit vulnerabilities that can lead to massive data losses, or other catastrophic consequences. From policymakers' and law enforcement's perspective, it is difficult to identify risk factors without firms' close cooperation, which may be impractical. Developing a tool that uses publicly available information to develop cyberrisk profiles can help policymakers and auditors prioritize their regulatory activities.

I utilize the fact that the Securities and Exchange Commission (SEC) requires numerous disclosures from publicly traded companies. In particular, every publicly traded company must provide a statement of its risk factors, and financial statements detailing the overall health of the company. The goal of these disclosures is to signal potentially relevant information to investors. In addition, companies must disclose financial information about their stock performance, tax liabilities, and assets to investors and regulators. The SEC is generally interested in harnessing its massive troves of data for Artificial Intelligence applications. I propose using machine learning and Natural Language Processing (NLP) techniques to train an algorithm that predicts future cybersecurity incidents based on firm-level the text of a company's filings. If successful, this could prove to be a valuable tool for regulators as they attempt to identify risky companies, and develop interventions

to prevent cybercriminals from exploiting those risks. Just as the FTC emerged in response to the growing problem of trusts, this study highlights the potential for the SEC to take on a similar role with regards to cybersecurity.

2 Law & Economics of Cyber Risk Disclosure

Cybersecurity regulation and auditing is an asymmetric information problem. Firms have private information about their cybersecurity posture, and this information is not available to regulators without intervention. Absent incentives to publicize this information, firms will prefer to keep this information private. In the cybersecurity context, the law has thus far addressed this problem by mandating disclosure of relevant cybersecurity risks alongside with other mandatory financial disclosures. Failure to adequately disclose this information can result in an investor lawsuit, thus imposing a cost on firms that fail to disclose risks, suffer an adverse cybersecurity event, and are subsequently sued.

From a regulator's perspective, obtaining information through these disclosures raises additional questions. Assuming the regulator is interested in obtaining the maximum amount of information about a firm's cyberrisk, it will craft disclosure requirements with an eye toward optimizing this quantity. These requirements' design is critical because firms are not likely to disclose relevant information unless there are other incentives to do so (such as signaling preparedness to regulators and investors). This simple model is the basis for the relationship between regulators and firms in a wide variety of contexts that involve audits, such as food safety inspections.

This simple model can be expanded by considering firms' own abilities to understand their cybersecurity postures. Although firms have private information about their policies, estimating cyberrisk requires domain expertise and involves uncertainty with regards to relative risk compared to similar firms. Again, because each firm's cybersecurity posture is private information, firms are unlikely to know what similarly situated firms are doing, and therefore cannot assess their own risk relative to their competitors.

There are several vendors who develop risk assessment tools that provide companies with cyberrrisk scores. For example, Security Scorecard uses a combination of information volunteered by a firm along with information scraped from a variety of security risk databases. It scores companies on ten different categories, and returns an A-F letter grade, along with access to a dashboard that helps companies pinpoint areas for improvement. Similarly, FICO offers a Cyber Risk Score service. Like its consumer credit scores, the scores seem to range between 300 and 850. Both of these services sell enterprise editions to companies and provide them with an comprehensible metric. These services therefore ameliorate the costs firms face with regards to processing their own information about their cyberrisks, and understanding their position relative to similar firms. These scores are also sold to insurers who underwrite cybersecurity incident policies, thus potentially solving the problem of distributing risk across similarly situated firms for rare events.

However, these scores remain private information for the firms in question, and therefore do not

solve the problem of regulators having less information about firms' relative cyberriskiness. A tool that provides information about firms' security postures to regulators would therefore help bridge this information gap. Moreover, while these services advertise that they use machine learning tools to generate their scores, and that the scores are a direct measure of riskiness with respect to suffering a breach, the validation strategies are unclear and not publicly available. Benchmarking firms' private assessments of their riskiness against real-world outcomes and making that information available to regulators would be helpful for refining and targeting regulatory efforts such as audits. Critically, creating a risk assessment based on public data gives regulators the ability to prioritize their decisionmaking even if firms do not volunteer to assess their own cybersecurity postures.

3 Literature Review

There are few studies that directly predict future cybersecurity incidents. In large part, the cybercrime literature is more concerned with causation rather than prediction. This is not unique to cybercrime however, as the social sciences traditionally focus on causation. However, techniques originating from data science open up opportunities to engage in useful prediction exercises as well.

Kleinberg et. al. argue this point in "Prediction Policy Problems." In this paper, the authors argue that machine learning techniques do not get adequate attention in the social sciences, and in economics in particular. They make the case that social scientists frequently miss interesting prediction questions because of the traditional focus on causal inference techniques. To illustrate, they use the toy example of a doctor deciding whether to perform a hip replacement on a patient. The catch is that the hip replacement is very painful in the short term, and would only improve the patient's qualify of life after about six months or so. Thus, it is only worthwhile to provide this treatment if the doctor can be reasonably sure that the patient will live at least that long. The prediction problem is trying to accurately predict whether a patient will live for six more months. If so, the hip replacement would be worthwhile. If not, there would be no need to put the patient through unnecessary pain (and expend the time and money on the needless procedure). To do this exercise, the decision maker does not need to know *why* the patient will live or not in the next six months, but rather simply whether or not they will. This insight is the key to understanding the motivation underlying this project (Kleinberg et al., 2015).

Susan Athey extends this discussion by discussing the intersection of machine learning, causal inference, and policy evaluation. In particular, she highlights the importance of rigorously mapping an algorithmic decision to an actual policy decision. While Kleinberg highlights a useful example of applying simple off-the-shelf methods to a problem, Athey argues that some understanding of the domain problem and causal mechanisms is still necessary for successfully implementing machine learning in policy. Pairing predictive decisions with techniques drawn from causal inference will help guide optimal policy decisionmaking (Athey, 2017)

Within the legal literature, Joshua Mitts makes a similar argument in "Predictive Regulation." He notes that regulatory agencies frequently design rules and interventions that respond to "crises"

or other events, but oftentimes, it would be better if these rules could be designed in a way that anticipated crises rather than correct them after the fact. He motivates this line of reasoning by pointing out that it is unlikely that the next financial crisis will be caused by subprime mortgage lending the way the 2008 crisis was. Regulators could avert the consequences of future crises by anticipating them and enacting relevant rules beforehand.

He argues that statistics provides many of the essential tools that can predict adverse events, and therefore enable policymakers to pro-actively intervene. Specifically, he demonstrates that natural language processing techniques could have flagged speculative language in the housing market before the 2008 financial crisis. He points to the potential for using these techniques across domain contexts could dramatically improve regulatory efforts (Mitts, 2014).

These papers provide a basis for exploring predictive cybersecurity policy. One need not understand the exact mechanics of the motivations of cybercrime to predict which companies are most at risk of suffering an attack. Instead, one must simply do a reasonably good job of predicting accurately, and therefore better informing decisions about where to target interventions.

The most direct study of predicting cybersecurity incidents is a paper from the University of Michigan entitled, "Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents." The authors in this study created an incidents database from a combination of the VERIS, Hackmageddon, and Web Hacking Incidents Database. These datasets constituted the outcome data, and they were joined with features drawn from each organization's cyber practices. These included features such as DNS misconfiguration, spam/phishing activity, etc. Overall, using a random forest algorithm, the authors report a high accuracy rate (90% True Positive, 10% False Positive) (Liu et al., 2015).

Aside from the cybercrime literature, there is a rich literature surrounding SEC disclosures. The theoretical foundations of corporate disclosure as a regulatory tool have been explored at length in the economics, business, and law literatures. These literatures ask questions about the optimal amount of disclosure to require, the incentives underlying honest and dishonest signaling in disclosure statements, and whether insiders use information to their advantage prior to a disclosure. These are all key questions that motivate the use of disclosure as a tool, and are particularly attuned to the SEC's disclosure requirements because of the high stakes involved with publicly traded firms, and the relatively consistent and stringent regulations placed on them.

Christian Leuz and Peter Wysocki provide a general overview of the various literatures in "Economic Consequences of Financial Reporting and Disclosure Regulation: A Review and Suggestions for Future Research." They identify a gap across the board, namely that the study of disclosure has largely focused on voluntary disclosures made by individual firms. In comparison, there is relatively little work done that studies the effects of mandatory disclosures, and how well those regulations achieve certain policy outcomes (Leuz and Wysocki, 2016). Various other studies that look at the effects of mandatory disclosure regulations generally focus on the effect of disclosure regulation and legislation on capital markets. The authors cite a number of studies that look at SEC regulations dating back to the 1930's that mainly look at how firms adjust behavior when a

regulatory regime is imposed, and how capital markets react when disclosures are made. However, these studies by and large do not make extensive use of natural language processing or qualitative analysis that examines the content of the disclosures themselves, and therefore miss key questions about the relationship between the regulations, disclosure content, and outcomes.

Kogan et. al. use the text of 10-K disclosures to predict stock price volatility. In particular, they used a tf-idf featurization and support vector regression technique to predict price volatility. They find that the text of 10-K disclosures provides substantial information to make predictions about historic price volatility (Kogan et al., 2009).

Otherwise, there is also a growing interest in the use of data science, machine learning, artificial intelligence, and other quantitative methods in the SEC. In a recent statement, Scott W. Bauguess, Acting Director and Acting Chief Economist of the SEC, articulated the SEC's goals in thinking about the rise of these methods. He emphasized that SEC regulators would benefit from being able to predict likely outcomes in a range of domains, and these tools provide unprecedented potential to do so. As part of its commitment to developing such technologies, the SEC makes troves of its own data and the raw text of disclosures available on its EDGAR interface (Bauguess, 2017).

The availability of these data has encouraged some preliminary work in implementing data science approaches to regulation. Joshua Mitts and colleagues wrote a piece entitled, "The 8-K Trading Gap" that looked at whether there was evidence of insider trading in the days preceding a damaging disclosure statement. Similarly in the cybersecurity context, Mitts and Eric Talley conducted a study that found evidence of insider trading prior to a cybersecurity breach disclosure (Mitts and Talley, 2018).

4 Data

4.1 Outcome Data

For outcome data, meaning reported data breaches and cybersecurity incidents, I combine several data sources that independently collect information about these events. In particular I use the Veris Community Database (VCDB) (which feeds into the Verizon Data Breach Investigations Report), the Privacy Rights Clearinghouse Chronology of Data Breaches Database, and the Identity Theft Resource Center. Each of these database maintainers collects different information and the nature of the incident. The most important distinction between these databases is the definition of breaches and incidents. Simply put, an incident can encompass a variety of events including loss of equipment, mismanagement of cybersecurity training, etc. Data breaches are one example of a cybersecurity incident. In general, companies do not always need to report incidents because they are not always material (in terms of securities regulation), but breaches are almost certainly material. Thus, the outcome data includes both breaches and material incidents, but it is important to note that these account for *reported* breaches and incidents. Because certain events, even material ones, may be unreported (and even undetected), focusing on reported breaches necessarily undercovers the universe of actual breaches and material incidents.

I use the Privacy Rights Clearinghouse (PRC) database as the baseline for data breaches and incidents, and augment the outcome data with breaches that are missing from PRC. I do this primarily because PRC collects information that is useful for sketching out the policy problem that may be missing from other databases, namely the type of incident, its location, and a description of the incidents. Figure 1 shows the number of breaches and incidents in the PRC dataset from 2010 onward. Note that these are breaches among publicly traded companies successfully matched in the dataset, as there are far more in the PRC database as a whole.

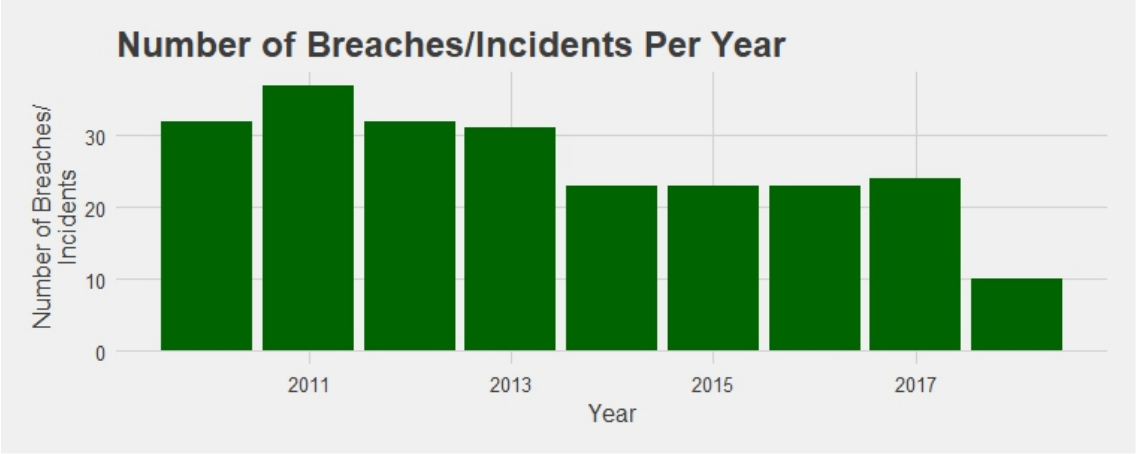

Figure 1: PRC Breaches Per Year

Notably, there are relatively few breaches among publicly traded companies in any given year. Some years have slightly over 30 breaches in this data, and closer to 25 in others. Compared to a universe of approximately 2000 companies [1], this makes breaches quite rare. In computer science terms, this is referred to an "imbalanced learning" problem because one class ("no breach") dominates in numbers over the other class ("breach").

Broken down by incident type, it is clear that the bulk of incidents is quite serious. In Figure 2, STAT refers to stationary computer loss, DISC to unintended disclosures, and PORT refers to portable device loss. Meanwhile HACK refers to outside hacking or malware infections, and INSD refers to a company insider intentionally breaching information. The number of incidents in the HACK category grows over time, while unintentional data losses become less frequent over time. This trend may suggest that companies are becoming more careful and better at preventing data losses that result from carelessness. On the other hand, outside attacks have grown over time, which can point to increased cybercriminal activity, or a substitution away from techniques like phishing toward more sophisticated techniques like malware.

Looking at the descriptions of the events paints a similar picture. Figure 3 shows a word

[1] According to the Wall St. Journal, there are approximately 3500 publicly traded companies in the U.S. However because of inconsistencies in how companies report their disclosures under different central index key numbers (ciks), matching disclosure text, financial information, and incident information is difficult. Future iterations of this work will work to complete the dataset used in this paper to include all companies across all U.S. stock exchanges. That being said, aside from a handful of notable exceptions (e.g. McDonald's), there are few breached firms that did not make it into the dataset.

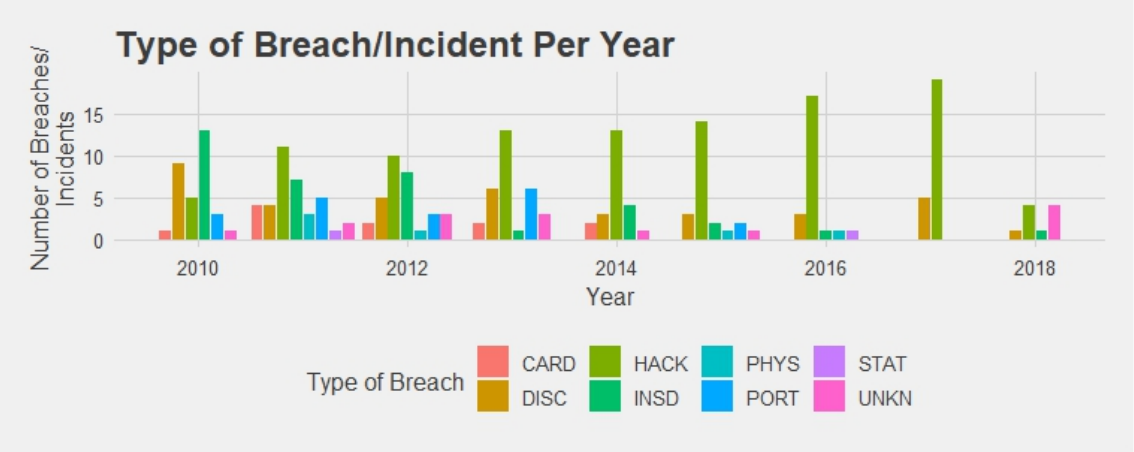

Figure 2: PRC Breaches Per Year and Type

cloud visualizing the most common words in the descriptions of the incidents. PRC writes these descriptions summarizing the description of the events from the source of the information about the breach (newspaper article, mandatory disclosure, etc.). Social security, credit card, bank, and email information are among the things talked about in these descriptions. These words give some idea of the sort of information that is most frequently compromised in these sorts of incidents among publicly traded companies. Geographically, incidents are concentrated in a handful of places. Firms in New York, New Jersey, and California make up the bulk of the outcome data. Given the prevalence of publicly traded companies in industries like finance and technology, this is unsurprising. Figure 4 shows the geographic spread.

Figure 3: Word Cloud of Description of Incidents

4.2 Firm-Level Data

In machine learning applications, text features tend to perform best when combined with non-text features. In this case, I collect firm-level data on each publicly traded company in my dataset.

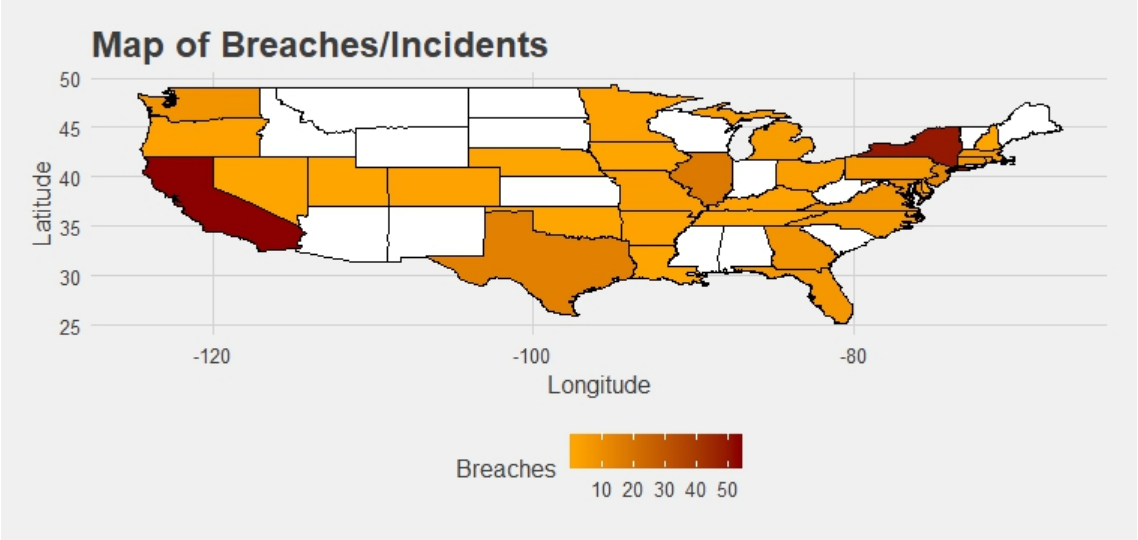

Figure 4: Map of Breaches and Incidents

These features are helpful primarily because they are already publicly available and easy to use, and therefore can provide a reasonable baseline for how regulators may try to predict data breaches without leveraging text information.

First, I extract industry codes and addresses. This information is helpful primarily because firms belonging to different industries will likely prepare for and respond to cybersecurity incidents in different ways. For example firms that handle health information are susceptible to stronger negative consequences stemming from breaches, and may be more likely to invest more in precaution as a result. One example of different incentives is that generally consumers do not have individual causes of action after the announcement of a breach, but generally do enjoy causes of action when protected health information is compromised. Industry codes are therefore potentially valuable information, and geographic information may also be relevant insofar as it may serve as a proxy for things like firm size, products, etc.

Industry codes are also interesting because different industries have varying cyberrisk profiles. Figure 5 shows the number of breached and non-breached observations among a subset of the most represented industries in the dataset. Some industries, such as real estate, are well-represented in the dataset, but suffer relatively few breaches or incidents. Figure 6 shows the ratio of breached observations relative to non-breached observations per industry. Although there are over 200 industries represented in the dataset, only approximately 40 suffer cybersecurity incidents at all. Notably, 50% of observations associated with the financial services industry also correspond to breaches. Telecommunications, software, and retail also have fairly high risk profiles.

I also incorporate firm level data from US Stocks Database maintained by the Center for Research in Security Prices. Table 1 summarizes the features drawn from the US Stocks Database. Critically, I avoid trying to predict how stocks may respond to cybersecurity incidents. Rather, I use stock volatility as a proxy for a firm's general riskiness, as measured by how investors respond in capital markets. (Kogan et al., 2009) already demonstrated that text analysis successfully

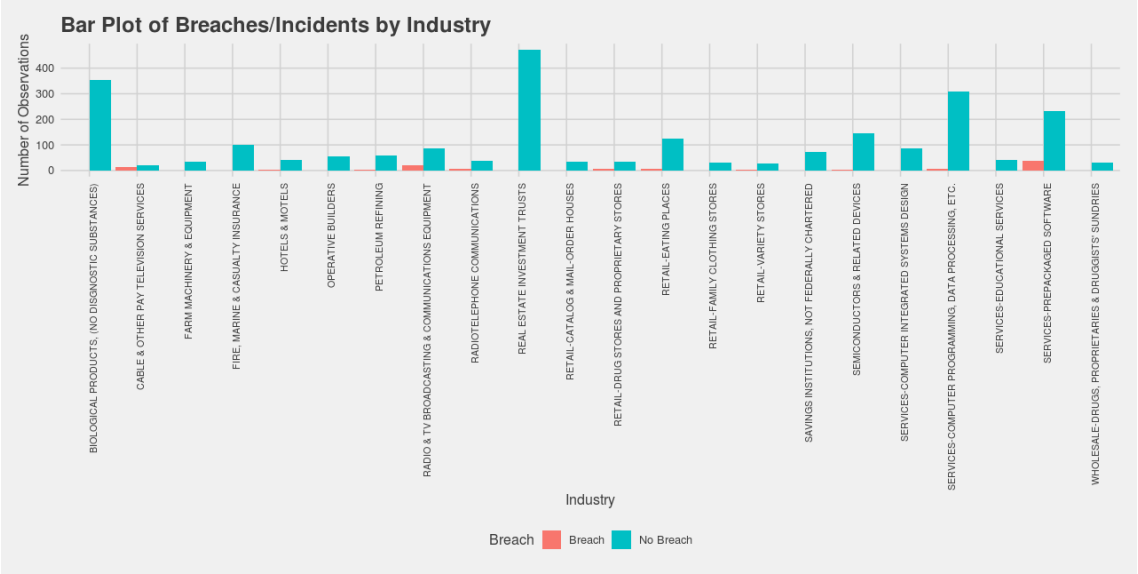

Figure 5: Industries with Breaches/Incidents

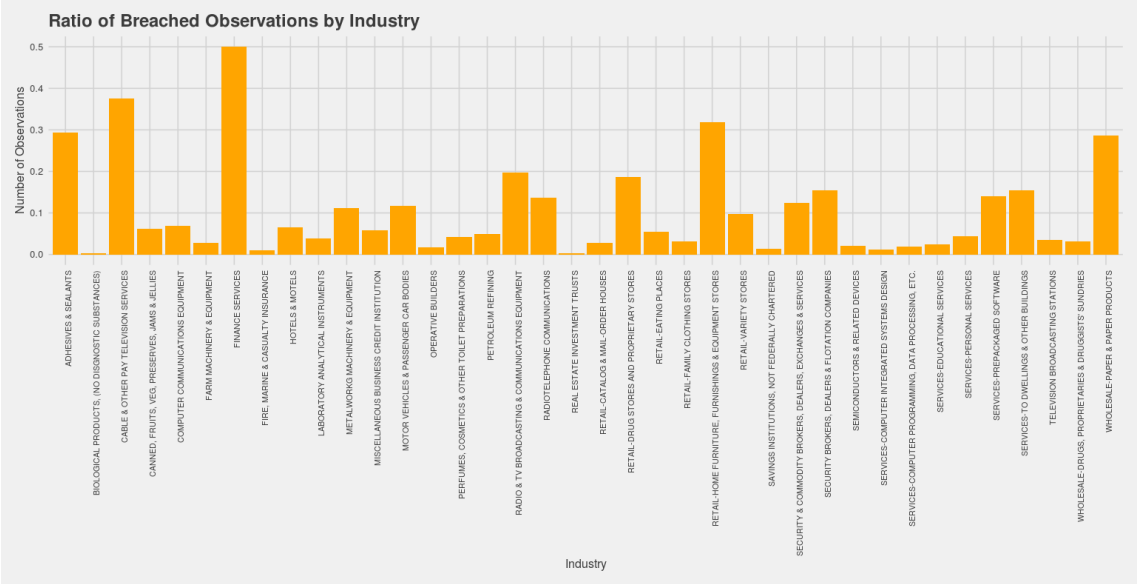

Figure 6: Ratio of Breaches/Incidents

predicts an asset's stability fairly well. The basic idea here is to take that measure of riskiness, and use it as a feature to predict cybersecurity riskiness.

4.3 Text Data

The text data source is the Securities and Exchange Commission's (SEC) datasets that collect companies' annual 10-K disclosures. In these 10-Ks, firms are required to disclose potential risk factors, including cybersecurity risks, to their investors. However, the SEC recognizes that companies need to manage the language in these disclosures so as to not create a roadmap for potential cybercriminals to exploit vulnerabilities.

Feature Name	Explanation
CURCD	Native Currency Code
TXDB	Deferred Taxes
TXDBCA	Deferred Tax Asset
TXDBCL	Deferred Tax Liability
TXDITC	Deferred Taxes and Investment Tax Credit
TXNDB	Net Deferred Tax Asset (Liability) - Total
TXNDBA	Net Deferred Tax Asset
TXNDBL	Net Deferred Tax Liability
TXNDBR	Deferred Tax Residual
TXP	Income Taxes Payable
CSHTR_C	Common Shares Traded - Annual - Calendar
DVPSP_C	Dividends Per Share
PRCC_C	Price Close - Annual - Calendar
PRCH_C	Price High - Annual - Calendar
PRCL_C	Price Low - Annual - Calendar
CSHTR_F	Common Shares Traded - Annual - Fiscal
MKVALT	Market Value - Total - Fiscal
ADDZIP	Zip Code
CITY	Headquarters City
State	Headquarters State
Industry Title	Standard Industry Code Industry

Table 1: Features Drawn from U.S. Stocks Database

4.3.1 Extracting Risk Disclosure Text

The most difficult data collection task is collecting all of the relevant SEC filings so that they can be matched to the outcome data. The SEC provides an online search tool (EDGAR) for looking up individual firms and their corresponding documents, but this does not lend itself to dataset construction.

Luckily, a number of open-source packages are available that aid with this task. In particular, I use the "edgar" and "edgarWebR" packages in the R computing environment. The edgar package provides a list of "Central Index Key (cik)" numbers that uniquely identify each publicly traded company. The edgarWebR package includes functions for looking up companies by their cik numbers, and extracting the raw text of their disclosures. A key feature here is that the package also includes a method for tagging parts of a disclosure, such that a user may tag all text that falls under the "Risk Disclosure" heading, which is always "Item 1A" on a 10-K disclosure form. Because some forms may be ill-formed, doing this computationally may not capture every relevant

aspect of every disclosure. However, it should be sufficient for most purposes.

After extracting the risk disclosure text, the next task is combining it with the outcome data and other features. Ultimately, the resulting dataset contains information about a firm's name, cik number, filing date, risk disclosure text, firm-level features drawn from the U.S. Stocks Database, and a logical indicator suggesting whether it was breached in the year following the publication of its risk disclosure. An example dataframe can be viewed here.

4.3.2 Exploratory Analysis

An example of the raw text of a disclosure can be seen here. This filing is from Apple's 10-K filing in 2011. Under its risk disclosure, it says the following about cybersecurity risks:

> The Company may be subject to breaches of its information technology systems, which could damage the Company's reputation, business partner and customer relationships, and access to online stores and services. Such breaches could subject the Company to significant reputational, financial, legal, and operational consequences.
>
> The Company's business requires it to use and store customer, employee, and business partner personally identifiable information ("PII"). This may include names, addresses, phone numbers, email addresses, contact preferences, tax identification numbers, and payment account information. Although malicious attacks to gain access to PII affect many companies across various industries, the Company may be at a relatively greater risk of being targeted because of its high profile and the amount of PII managed.
>
> The Company requires user names and passwords in order to access its information technology systems. The Company also uses encryption and authentication technologies to secure the transmission and storage of data. These security measures may be compromised as a result of third-party security breaches, employee error, malfeasance, faulty password management, or other irregularity, and result in persons obtaining unauthorized access to Company data or accounts. Third parties may attempt to fraudulently induce employees or customers into disclosing user names, passwords or other sensitive information, which may in turn be used to access the Company's information technology systems. To help protect customers and the Company, the Company monitors accounts and systems for unusual activity and may freeze accounts under suspicious circumstances, which may result in the delay or loss of customer orders.
>
> The Company devotes significant resources to network security, data encryption, and other security measures to protect its systems and data, but these security measures cannot provide absolute security. The Company may experience a breach of its systems and may be unable to protect sensitive data. Moreover, if a computer security breach affects the Company's systems or results in the unauthorized release of PII, the Company's reputation and brand could be materially damaged and use of the Company's products and services could decrease. The Company would also be exposed to a risk of

loss or litigation and possible liability, which could result in a material adverse effect on the Company's business, results of operations and financial condition."

This disclosure represents just one case, and Apple may be more conscientious than most companies. That being said, this type of language reflects the sort of text that might distinguish various cybersecurity practices. If there are patterns in the language, details, and other information presented in cybersecurity risk disclosures, this may emerge through natural language processing.

More generally, we can see general patterns in the way that companies talk about their risks. Figure 7 shows a topic model for two topics, trained on the text of the risk disclosures. These topics give a sense of the sorts of terms that are likely to appear together in a disclosure. Specifically, the concepts of "risk," "price," and "adverse" seem to come up, which should not be surprising given the nature of section 1A.

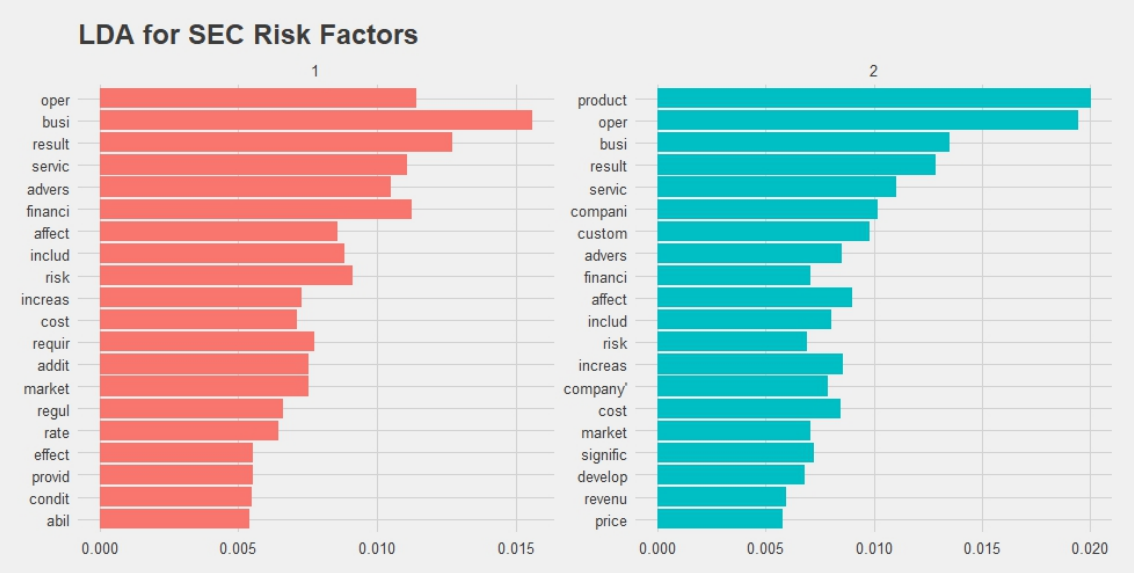

Figure 7: Latent Dirichlet Allocation for 2 topics

4.4 Feature Engineering

In addition to firm-level and textual data, I also conduct feature engineering to manually create some features that may be helpful for prediction purposes. From the firm-level data, I calculate the difference between high and low stock prices for the year to reflect stock volatility. I also make a logical indicator for companies that experiences breaches or incidents in previous years. Finally, I calculate the ratio of breached observations to unbreached observations within an industry.

I also used keyword searches of the disclosure text to create features that mapped to the SEC's interpretative guidance. Some examples of manually created features and the associated keywords can be seen in Table 2. Further feature engineering would use more sophisticated methods to pick up on the concepts underlying the SEC guidance, but keywords are a first attempt to see how basic models would do. Concretely, the SEC interpretative guidelines look at the following elements:

- Occurrence of prior cybersecurity incidents
- Probability of the occurrence and potential magnitude of cybersecurity incidents
- Preventative actions taken to reduce cybersecurity risks and associated costs
- Aspects of business that give rise to material cybersecurity risks
- Costs associated with maintaining cybersecurity protections
- Potential for reputational harm
- Existing or pending laws that might affect cybersecurity risk
- Litigation, regulatory investigation, and remeditation costs associated with cybersecurity incidents

Feature Name	**Key Words**
Probability of Occurrence	Cyberattack, Previous Incident
Preventative Actions	IT Security, Encryption, Cybersecurity Awareness Training
Aspects of Business	Personal Data, PII, PHI, Password
Reputational Harm	Harm to Our Reputation, Reputational Harm
Existing Laws and Regulation	Produce User Data, User Data Requests, Government Requests for User Data

Table 2: Features Engineered from SEC Interpretative Guidance

5 Policy Setup & Exploratory Analysis

In this section, I sketch out the decisionmaking process for SEC cybersecurity audits. I describe the substance of cybersecurity audits, as well as trends in how many have been conducted over the last few years. I then provide a simulation of how well randomly choosing firms to audit does at predicting future breaches. I then provide a simple model that uses a logistic regression to estimate the likelihood of a breach.

5.1 Background

The SEC is increasingly paying attention to cybersecurity risks and is taking active steps to safeguard investors. In 2017, the SEC established a Cyber Unit in its Division of Enforcement. According to the SEC's website, the "Cyber Unit focuses on violations involving digital assets, initial coin offerings and cryptocurrencies; cybersecurity controls at regulated entities; issuer disclosures of cybersecurity incidents and risks; trading on the basis of hacked nonpublic information; and cyber-related manipulations, such as brokerage account takeovers and market manipulations using electronic and social media platforms." Most of the enforcement actions brought so far deal with

initial coin offerings, but the SEC also pursues actions related to failure to adequately disclose material events and cyberrisks.

In 2017, the Office of Compliance Inspections and Examinations (OCIE) conducted a pilot program where it audited the cybersecurity policies and practices of 75 publicly traded firms. It found that while firms generally had written policies in place about how they should deal with cyberrisk and adverse events, oftentimes these written explanations were too vague to provide helpful guidance to employees. Moreover, it was not always clear that firms actually implemented some of their written policies, such as requiring and monitoring cybersecurity training. In general, the SEC is expanding its auditing and enforcement efforts, as the number of firms subject to some kind of audit (not just cybersecurity) increased from 8% to 13% from 2013 to 2018. As part of this general expansion, the SEC is paying particular attention to cybersecurity concerns.

5.2 SEC Cybersecurity Audits

In 2015, the SEC launched its Cybersecuriry Examination Initiative. With this notice, the SEC outlined the general procedure for its cybersecurity audits, and what minimum standard firms are expected to uphold. The specific areas that SEC examiners focus on are:

- Governance and Risk Assessment
- Access Rights and Controls
- Data Loss Prevention
- Vendor Management
- Training
- Incident Response

In these audits, the examiners look at both a company's written policies, as well as their actual practices. There is now a cottage industry surrounding preparedness for these cybersecurity audits. One source suggests that an audit may take about six days, and requires three SEC auditors (one of whom specializes in cybersecurity audits).

5.3 Metrics

Before providing baseline simulations to motivate the core policy problem, I define basic metrics for evaluating the efficacy of cybersecurity audits. Simply put, the prediction task here is predicting whether a firm will suffer a cybersecurity breach or incident. There are various ways to define whether a prediction task is working well. In this case, the task is predicting "breach" or "no breach," with "breach" being the "positive" class. Some foundational building blocks to think about predictions in this case include:

- **True Positives (TP)**: Predictions where the model accurately predicts the positive class. In this case, these are instances when a model predicts a "breach" and there was indeed a breach.
- **False Positives (FP)**: Predictions where the model erroneously predicts the positive class. In this case, these are instances when a model predicts "breach" when there was no breach.
- **True Negatives (TN)**: Predictions where the model accurately predicts the negative class. In this case, these are instances when a model predicts "no breach" and there was no breach.
- **False Negatives (FN)**: Predictions where the model erroneously predicts the negative class. In this case, these are instances when a model predicts "no breach" when there was actually a breach.

In this context, true positives and false negatives are the most consequential metrics. Successfully predicting a true positive indicates that the model found an ideal candidate for an audit, while predicting a false negative (failing to detect a breach) implies a situation where an audit may have helped but the model failed to direct the intervention toward that firm. False positives imply that the model would have a firm audited that may not have needed it, and while this imposes costs on the agency, is not as consequential as a false negative. Meanwhile, true negatives are trivial to predict in this context because relatively few firms are breached in any given year.

Delving deeper, these metrics can be combined in useful ways.

$$\textbf{Accuracy} : \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Accuracy is essentially a measure of how many times the model was correct in its predictions in either direction, divided by the total number of predictions it made.

$$\textbf{Recall} : \frac{TP}{TP + FN}$$

Recall is a measure of successful the algorithm was at detecting instances of the positive class. In this case, the ratio is an expression of what fraction of the actual breaches the algorithm successfully predicts.

$$\textbf{Precision} : \frac{TP}{TP + FP}$$

Precision is a measure of how successful a model at filtering out noisy predictions. Put differently, it is a statement of what fraction of all the predictions of the positive class were actually in the positive class. In this case, it is saying of all the firms that the model predicted would be breached, how many were actually breached.

5.4 Random Audits

In 2015, the SEC began its cybersecurity auditing program. That year, the SEC conducted 75 audits. I simulate these audits to provide a baseline for an algorithm to improve upon. To do this,

I simulate randomly choosing 75 firms to audit, and plot the distributions of how well these audits predict eventual breaches. I ran 100,000 simulations of picking 75 firms to audit at random, and then plotted distributions for true positives, recall, and precision. To accomplish this, I looked at breaches for the 2015-2016 fiscal year, where there were 10 breaches.

Figure 8 shows the distribution of true positives across these simulations. Across 100,000 simulations, the modal outcome is to successfully detect 0 breaches in advance. In the tail of the distribution, randomly auditing may pick up on one or two eventual breaches, but hardly ever exceeds these figures. Similarly, recall follows the same pattern, as seen in Figure 9.

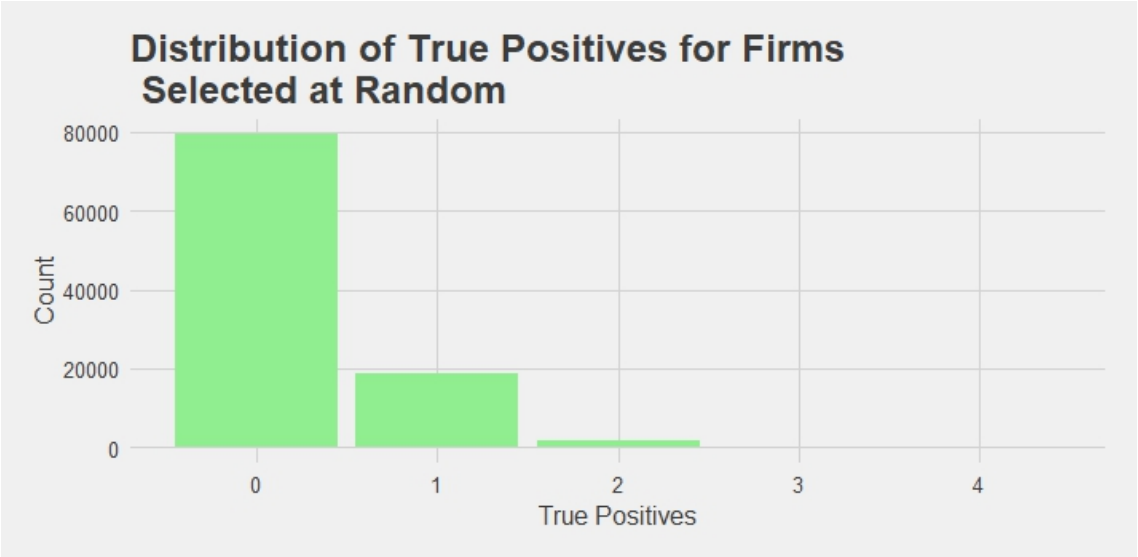

Figure 8: Distribution of True Positives in Random Audits

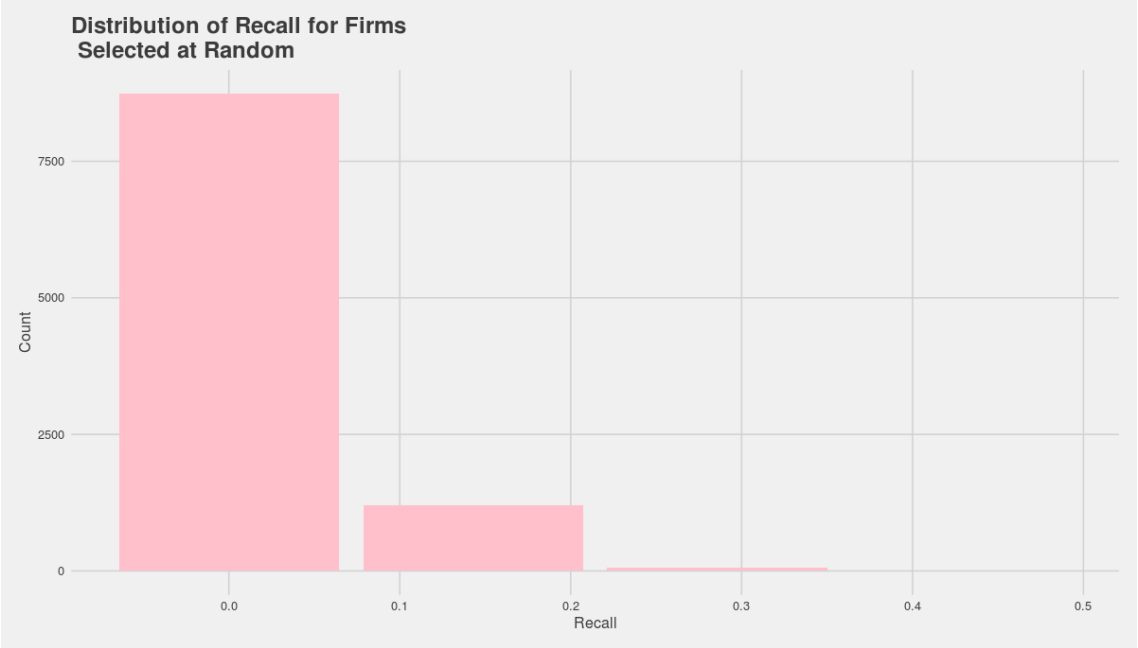

Figure 9: Distribution of Recalls in Random Audits

Even if an algorithm could not improve on true positive and recall measures, there is substantial room to improve on precision. Figure 10 shows the distribution of precision across simulations of random audits. Even in cases where a random audit successfully predicts a breach, the precision score lies somewhere between .015 and .025. The takeaway here is that of the 75 audits conducted, about 73-74 are potentially wasted. Any improvement over this precision score would make these audits more efficient.

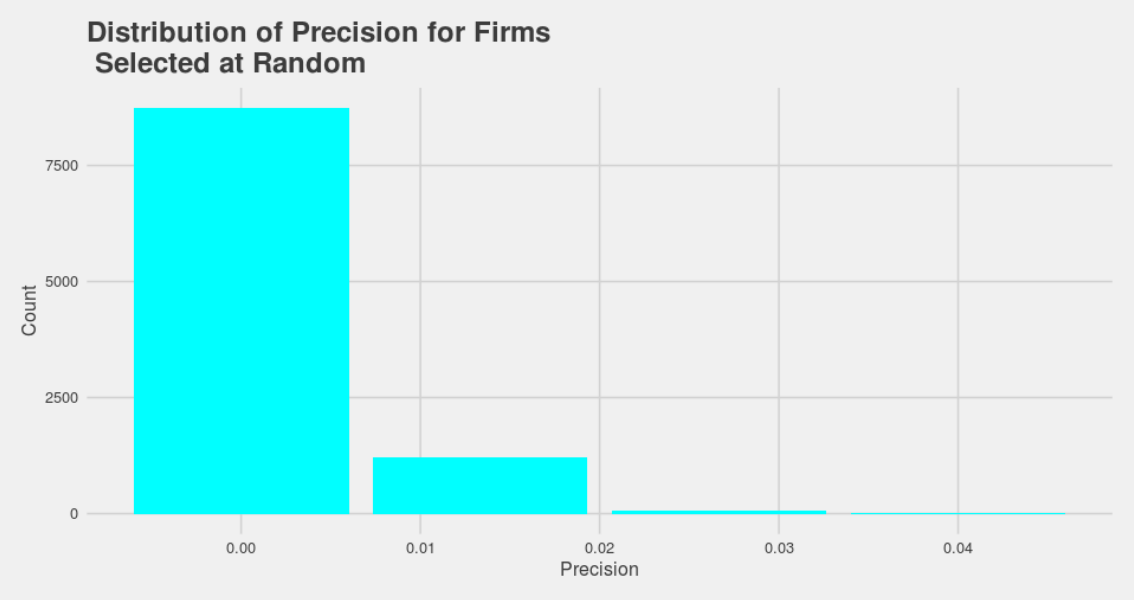

Figure 10: Distribution of Precisions in Random Audits

5.5 Logistic Regression

Next, I train a logistic regression to simulate how well a simple algorithm performs on this prediction task. Using the same subset as I did with the random audits, I train the logistic regression on the disclosures. In this case, I featurize the text of the disclosures using the term frequency-inverse document frequency (tf-idf) technique. The simplest natural language processing (NLP) model that could be used is the "bag of words" model where the columns in the dataset correspond to counts of how many times a given word appears in a document. In this context, a document is a 10-K disclosure for a particular company and year. Instead of using a bag of words, tf-idf takes the number of times a term appears in a given document (term frequency), and then multiplies that by the inverse of the number of documents that the term appears in (idf). The basic intuition here is that more weight is given the more times a term appears within a document, but then weight is decreased the more common a term is across documents. Thus, tf-idf does well classifying documents where individual documents have key terms that do not appear elsewhere in the corpus.

Table 3 shows a confusion matrix that shows how well a logistic regression does with tf-idf weighting at predicting outcomes for the 2015-2016 fiscal year. A confusion matrix is a useful

tool for visualizing how well an algorithm did at a classification task, and common metrics like accuracy, recall, and precision are easily derived from it. Of the 4209 companies in this dataset, 5 suffered breaches. The model predicted 4 companies would suffer breaches, but none of these predictions overlapped with actual breaches. The model did successfully predict almost every case of "no breach," but this is of little value given the severe imbalance in the dataset.

	observed no breach	observed breach
predict no breach	4200	5
predict breach	4	0

Table 3: Logistic Regression Confusion Matrix

In this case, the simple model does even worse than random auditing. While a combination of logistic regression and tf-idf has advantages with regard to transparency and interpretability, more complex models are likely to do better in this case. Given that this logistic regression had 0 true positives, it similarly had a recall and precision of 0.

The low baselines implied by both random auditing and logistic regression motivate the possibility for exploring other methods that can enhance successful prediction of cybersecurity incidents. The poor performance of a simple logistic regression may also point to why the SEC and other regulatory agencies have thus far been slow to adopt algorithmic approaches to prediction policy problems.

6 Methodology

6.1 Featurizing Text

To featurize the text (turn text into quantitative information), I use word2vec. Word2vec is a set of popular word embedding models first introduced by Mikolov et. al. (Mikolov et al., 2013). Word2vec is a "word embedding" technique, meaning it converts words into numerical vectors, and puts substantively similar words into vectors that are close together. Specifically, I use document-averaged word embeddings from word2vec to transform the raw text of annual disclosures into quantitative features.

6.1.1 Frequency-Based Featurization

The simplest model for featurizing text would be the "bag-of-words" approach. A bag-of-words is a frequency-based scheme that essentially counts how many times a word appears in a document and creates a feature for that count. One popular extension of the bag-of-words technique is the "term frequency-inverse document frequency" approach which counts the number of times a word appears in a document, but divides that figure by the number of times that word appears across a corpus. Thus, words that are unique to a document will get higher weights than words that appear frequently across documents.

Frequency-based approaches are useful because they are easy to implement and interpretable. Tf-idf in particular can be quite powerful when dealing with a classification task where there are words that are good as discriminating between class labels. For instance, in e-mail spam detection, certain key words show up in spam e-mails that rarely show up in legitimate ones. A model trained on tf-idf or bag-of-words features would be able flag spam simply by looking at whether words associated with spam class labels appear in a document. The disadvantage of these approaches is that in more complicated classification tasks, frequency based embeddings may sacrifice too much of the substantive meaning of words to be useful predictors. Whereas in the spam example there is a clear link between the presence of some words and the outcome label, this relationship is not always so strong. In the context of this study, word frequencies likely do not so neatly map into the outcome of a firm being breached, thus warranting considering more information.

6.1.2 One Hot Encoding

One way to capture more of a word's meaning in context is the one-hot encoding approach. A one-hot encoder essentially takes a collection of words (sentence, paragraph, or document), and creates logical indicators for whether words in the corpus appear in that collection. For example, if we had five words in a feature space, "I," "love," "data," "is," and "cool" then the following sentences would be encoded as follows:[2]

	I	love	data	is	cool
I love data	1	1	1	0	0
love is cool	0	1	0	1	1
data is cool	0	0	1	1	1

Table 4: One-Hot Encoder

This featurization approach is useful primarily because it encodes sentence-level (or paragraph/document) information in a numerical vector. [3] By representing sentences as vectors, more information about the distance between sentences is available to the analyst. However, the one-hot encoding still does not understand the meaning of the sentences. Although "love is cool" is close to "data is cool" in vector space, these vector representations still depend on the appearance of certain words, rather than their actual substantive meaning. Moreover, in this application I am looking at entire documents. The feature space quickly becomes high-dimensional when one-hot encoding thousands of long documents, which increases computational complexity.

[2]This example is borrowed from: https://towardsdatascience.com/an-intuitive-explanation-of-word2vec-208bed0a0599

[3]A "vector" in this case should be understand in its linear algebra context. A vector represents an object with a magnitude and direction (for instance, the acceleration of an object), and vectors can be operated on in a vector space, which is a collection of vectors. In this case, the vectors encode information about a sentence, and situates each sentence in a vector space relative to other sentences.

6.1.3 Word2Vec

Word2Vec is a word embedding technique that uses a prediction-based approach to creating word vectors. The training process for word2vec involves predicting words based on their surrounding context. [4] Using the context surrounding a word as input, this context is passed through a neural network to produce vectors with probabilities to predict target words. The algorithm then uses a technique called backpropogation to adjust the weights it assigns to these vectors until it minimizes the loss function (or more simply, until it does as well as it can at predicting words). This process then outputs vector representations for each word, and words with similar contexts will have vectors that are closer together in vector space. The canonical example of the advantage of this approach is encapsulated in this relationship:

$$\vec{king} - \vec{man} + \vec{woman} = \vec{queen}$$

Taking the vector for king, subtracting the vector for man, and adding the vector for woman yields a vector that is very close to the vector for queen. Thus, the word2vec representations are able to capture the idea that the concept of a queen is similar to king, except for a difference in gender. Thus, these word2vec vectors are able to capture more contextual meaning than word frequency or order.

I use word2vec and update the vectors with the text of the SEC disclosures. I then take these vectors, and average them across documents. Doing so creates a document-level vector that is built upon the tuned word vectors. These document-level vectors then become features in the downstream classification task, which is predicting firms that are likely to suffer cybersecurity incidents.

6.2 Modeling

6.3 Constituent Models

I use several constituent models before fitting an ensemble model. Importantly, each of these models is well-suited to classification tasks, though some can be used for regression as well. In machine learning terms, a classification task is distinguished from a regression task by the nature of the target variable (the variable that we are trying to predict). Classification is predicting which class label an observation belongs to. Binary classification predicts a target that can take one of two class labels, whereas multi-class classification predicts targets with many labels. In contrast, regression predicts continuous target variables. In this case, predicting whether a firm is breached or not is a binary task.

[4] This prediction can either use Continuous Bag-of-Words (predicting a target word from surrounding words) or skipgram (predicting surrounding words from a target word). CBOW does better with larger datasets and common words, whereas skipgram is better for smaller datasets and rare words. See Figure 11 for an illustration.

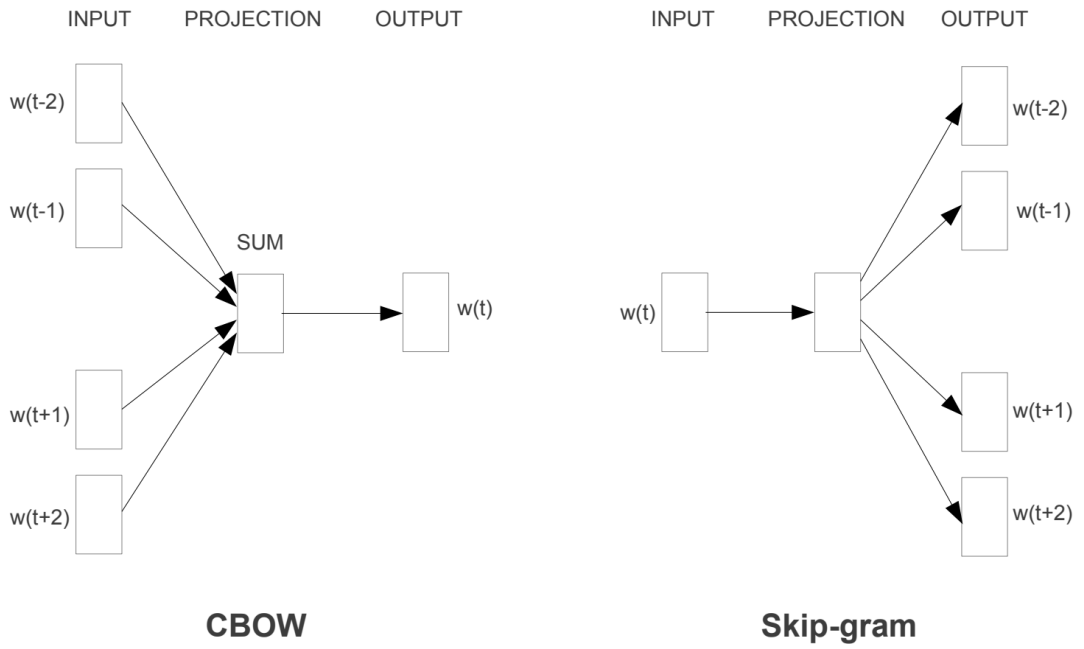

Figure 11: Illustration of Continuous Bag-of-Words and Skipgram, taken from Mikolov et. al. 2013

6.3.1 Logistic Regression

Logistic regression (logit) is a common algorithm in the social sciences, and is especially popular for binary classification tasks. Most social science applications of logistic regression report coefficients on the features (independent variables or covariates in social science language) for causal estimates. These coefficients are generally reported as log-odds, though sometimes are exponented to odds ratios. Critically, in a prediction context, the coefficients are not the object of interest for analysts. Rather, only the predicted probabilities for the target in the test set are relevant for the analysis. In a prediction setting, the causal interpretation of various coefficients is not especially relevant because a policymaker does not need to understand the precise relationship between the outcome and a feature to make a decision.

6.3.2 Poisson Regression

Poisson regression is a generalized linear model that is popular for modeling count data. Generally, poisson models are not used for binary outcome data, but I use one here because of poisson's strength in modeling rare events. In this case, poisson is akin to using a linear probability model. Essentially, these approaches use a linear model to estimate a binary outcome. The main disadvantage of these approaches is that without restrictions, it is possible to predict values outside the range [0,1], which would be invalid for a truly binary outcome.

6.3.3 Classification Tree

Decision tree learning is a machine learning approach that predicts a target value through a series of decision rules. Trees can be used for both classification and regression problems. The basic idea behind a decision tree is that learns the relationships between features and targets by growing a tree that encapsulates various decision rules. The tree starts at an initial node, and then makes a split into two new nodes based on some decision rule. At each of these nodes, the tree then splits again based on a new rule. This process iterates until the tree cannot make any more splits. A frequently used example to illustrate this concept is a classification tree that predicts whether a passenger on the titanic would survive given the rules for who was allowed to board a lifeboat (See Figure 12).

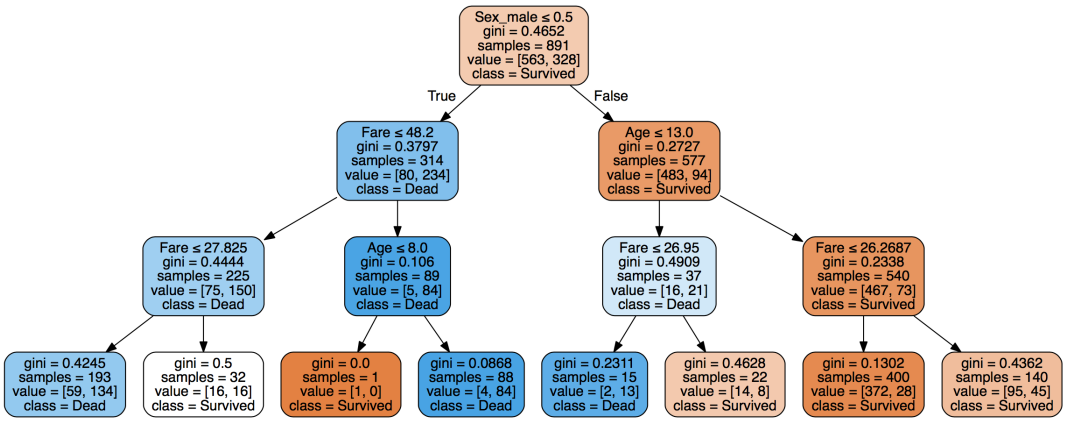

Figure 12: Titanic Survival Classification Tree

6.3.4 Random Forest, Gradient Boosting Classifier, and Adaptive Boosting

Classification trees have a few drawbacks, however. Without pruning (reducing the depth of a tree), trees tend to overfit the data, thus achieving poor performance out-of-sample. Trees also initialize from a randomly chosen feature, and make probablistic splits. Thus, any given tree may be overfit to idiosyncrasies in that particular random sample. To address these problems, classification trees are frequently combined in "ensemble" methods.

One approach to solving these problems is using a "bagging" technique such as a random forest. A random forest grows many classification trees in parallel, and then has each tree vote for the outcome. The prediction with the majority vote is the final prediction for the random forest. Random forests are popular because they reduce the tendency of single trees to overfit, and can be trained quickly with parallel processing.

Another ensembling approach for trees are "boosting" algorithms. Whereas bagging grows trees in parallel, boosting instead iteratively combines weak classifiers (classifiers that do slightly better than a coin toss at predicting an outcome) to create a strong classifier (a classifier that has close to 0 error). Boosting takes longer to train than bagging because it is iterative, but has the advantage

of having each sequential model learn from the mistakes of the previous models. In this case, I use gradient boosting and adaptive boosting, which primarily differ in how they combine weak learners. Gradient boosting learns from the errors (pseudo-residuals) in the previous iteration of the algorithm. Adaptive boosting learns by upweighting data points that were incorrectly classified in the previous iteration, thus forcing the algorithm to learn how to deal with more difficult decisions.

Ensembling trees is especially attractive in an imbalanced dataset setting. In this case, the cases of "no breach" far outnumber the "breach" observations. Ensembles are better at predicting minority class observations because they reduce noise from overfitting, and are built in a way that focuses on harder cases.

6.3.5 Soft Voting Ensemble Learner

Finally, I take all of the constituent algorithms, and combine them into a soft-voting ensemble classifier. Using the predicted probabilities from each model, the ensemble takes the average of these probabilities and makes a decision based on that average. This approach can be contrasted with hard-voting classifiers where each model takes a vote, and the majority vote is the ensemble's decision. The soft voting ensemble takes advantage of the fact that each of these models outputs predicted probabilities, and combines them into a meta-learner.

Ensemble classification is helpful primarily because it ameleriorates idiosyncrasies that may plague any individual model. Moreover, knowing the "correct" model a priori is impossible, and ensembles help approximate the best possible model by averaging constituent models. Ensembles take advantage of the fact that if each model is more likely than not to make the correct prediction, combing their predictions will boost the accuracy because it is less likely that idiosyncratic errors in one model will turn into incorrect predictions.

6.3.6 Temporal Cross-Validation

A potential problem in building machine learning models on temporal data is the tendency for future information to leak into the training process. In typical machine learning modeling, the analyst splits the data into train and test sets (sometimes adding a "validation" set as well). The train/test split is done at random in most applications, and the model is then trained on the training data, and its predictions are compared to the true observations in the test data. However, this framework quickly breaks down with temporal data. If the splits are done randomly with temporal data, the machine learning algorithm learns patterns from a future time period, and its performance will be artificially boosted when it is tested on data from a previous time period. For instance, imagine if the training set randomly included the Target 2013 breach outcome in its training data, and the test set included the 2011 and 2012 financial disclosures timeframes. When testing the algorithm, it will almost assuredly predict a breach because it borrows information from a future year. This would make the algorithm seem accurate, but would not reflect actual deployment conditions, as a regulator will not have advance notice of a breach (indeed, such information would obviate the need for an algorithmic approach).

Instead, I utilize a "temporal cross-validation" approach. The intuition here is that the model is built sequentially so that it never borrows information from the future. Using a k-fold approach, each fold will represent a sequential year. Each successive fold is a superset of the previous fold, thus ensuring that only past information is used. For instance, for each entity, we might aggregate features in 2011 and 2012, train on outcome data from 2013, and then validate/test on outcome data from 2014. (for Data Science and at the University of Chicago, 0). Figure 13 illustrates the basic logic of temporal cross validation.

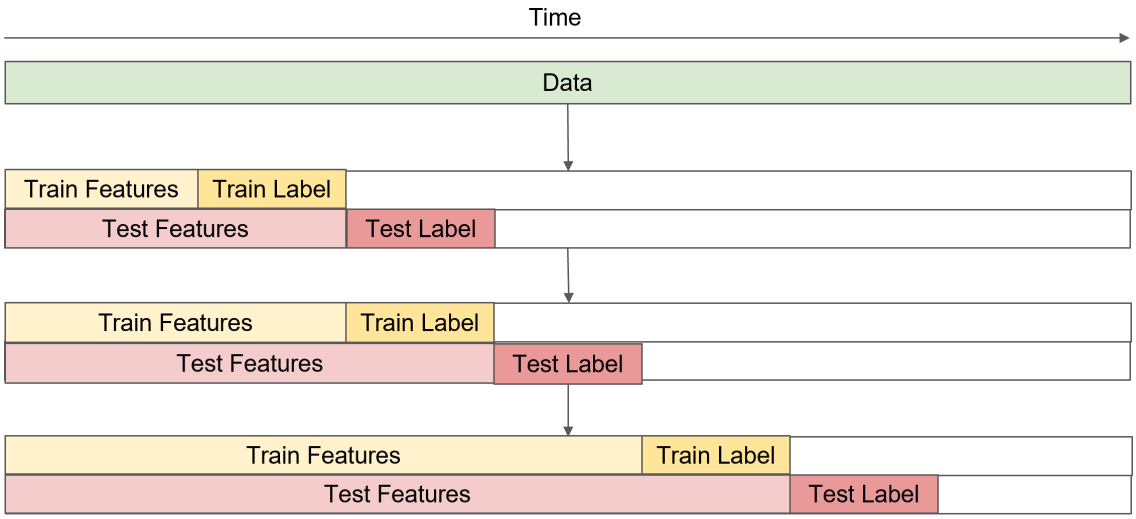

Figure 13: Illustration of Temporal Cross-Validation

6.3.7 Over and Undersampling

The major problem with predicting cybersecurity incidents is that although they are costly, they are relatively rare. In machine learning terms, this translates to an imbalanced learning problem. Essentially, instances of the majority class ("no breach") vastly outnumber instances instance of the minority class ("breach"). Thus, if an algorithm was trained to optimize only for accuracy, it would do quite well by simply picking the majority class every time. From a policy perspective, optimizing for accuracy alone is not always fruitful because regulators are oftentimes concerned with detecting and preventing rare but significant events.

One way to overcome this problem is to utilize over- and under-sampling techniques. Oversampling takes instances of the minority class and upsamples them in the training process, whereas undersampling takes instances of the majority class and downsamples them. Oversampling comes at the cost of potentially overlearning idiosyncrasies in minority class, and thus generalizing poorly. Undersampling comes at the cost of throwing away potentially relevant and useful information, thus reducing the algorithm's overall accuracy.

In this application, I combine over- and under-sampling together. Combining both helps capture some of the benefits of each, while ameliorating the disadvantages of each. In future iterations, I may look to other techniques such as Synthetic Minority Oversampling Technique (SMOTE) and

Random Oversampling Examples (ROSE) instead of a simple oversample. SMOTE in particular may yield benefits as it avoids some of the overfitting problems of simple oversampling.

7 Results

Although more work is necessary before deploying a model in this context, early results are promising. Compared to random audits or no audits, algorithmic predictions more successfully target risky companies and industries. In this section, I present the results of various model configurations. I present a baseline model with just firm-level information, a model trained only on text, a model that combines both firm-level information and text, and a final model that selects the most predicitve firm-level features, and discards unimportant features. In general, combining firm-level data and text features works best for prediction, and reducing model complexity aids with improving precision.

7.1 Baseline Results

The first model I present is a baseline model that uses only firm level features. These include the features described in Section 4.2. Figure 14 shows the results for this baseline model without any additional text features. I use logistic regression, poisson regression, classification tree, random forest, gradient boosting classifier, adaptive boosting methods. Although logistic regression performs quite well on recall (the ratio of firms predicted to be breached over the firms that were actually breached), this performance comes at the expense of accuracy (ratio of correct predictions to incorrect predictions) and precision (ratio of correct predictions to correct plus incorrect predictions). Essentially, the logit model here too aggressively guesses the positive class ("breach") in this case. The tree-based methods trade off some of this recall for more precision, though still are not as precise as we might hope for in a policy application. At best, the tree-based methods achieve around a .1 precision. While recall is more important in this application, too low a precision score implies that the SEC would erroneously flag too many audit candidates. Given limited resources, enough to conduct about 75 audits per year, flagging too many candidates potentially misses some risky targets.

Figure 15 shows feature importances from the random forest model. As suggested by the exploratory, industry riskiness is an important feature for predicting breaches. Proxies for firm size such as market value and tax liabilities are somewhat predictive, as are measures of stock volatility. Notably, only a few indicators for industry (software, retail) and geography (New York and Chicago) are predictive, with other dummy variables for these values taking on 0 or very low feature importance.

7.2 Text Only Results

Next, I show models with only text features in Figure 16. These models exclude any other firm-level information, as well as the manual feature-engineering of key terms corresponding to SEC

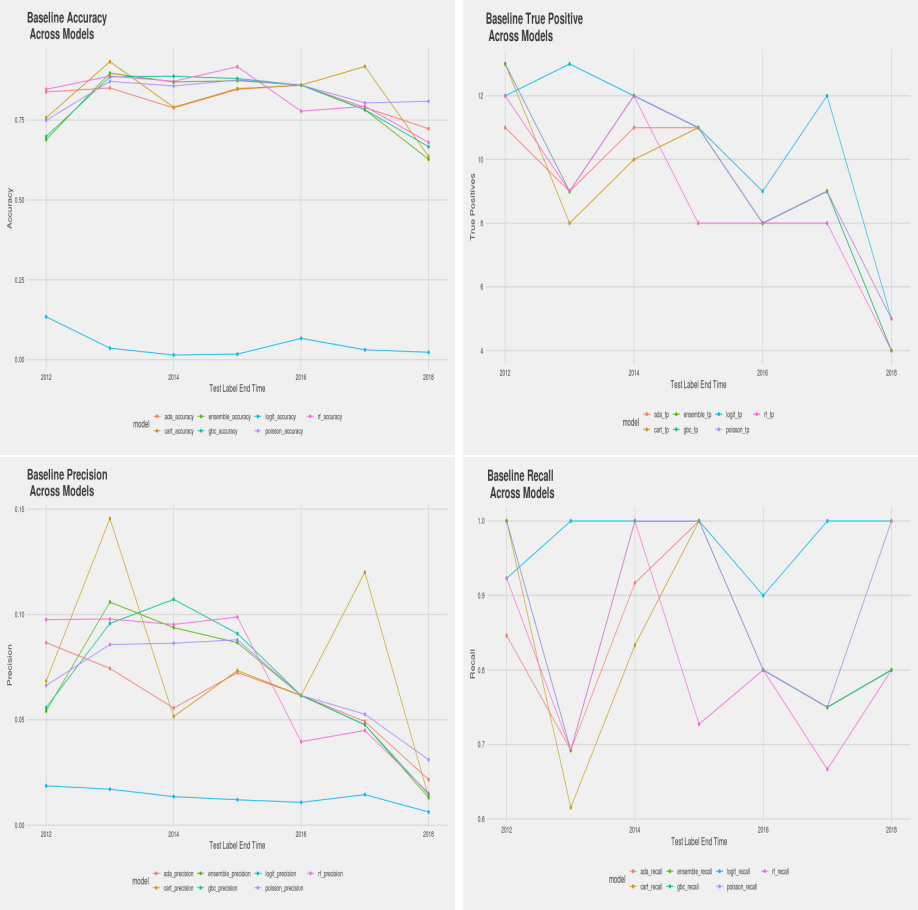

Figure 14: Baseline Results

interpretive guidance for describing cybersecurity risks. These results underperform the baseline models, regardless of the particular model chosen. In general, this result is not surprising. That being said, the text alone does seem to be somewhat informative.

7.3 Baseline + Text Results

Combining the baseline features with text features performs similarly to the baseline alone. While some models make some gains on recall, this may just be noise. Precision also seems to be a bit lower across models. Figure 17 illustrates these results. Again, these results are driven by including all possible features in the model, potentially leading to overfitting.

7.4 Selected Features Results

For the final models, I remove the unimportant firm level features from the baseline features and retrain each model. Removing these features and retraining the models considerably improves precision, though at the cost of some recall. Figure 18 illustrates this tradeoff. Looking at random forest, gradient boosted classifier, and adaptive boost, precision improves to about .4 in most years, though recall drops to about .5. In this context, this tradeoff is probably worthwhile as the higher

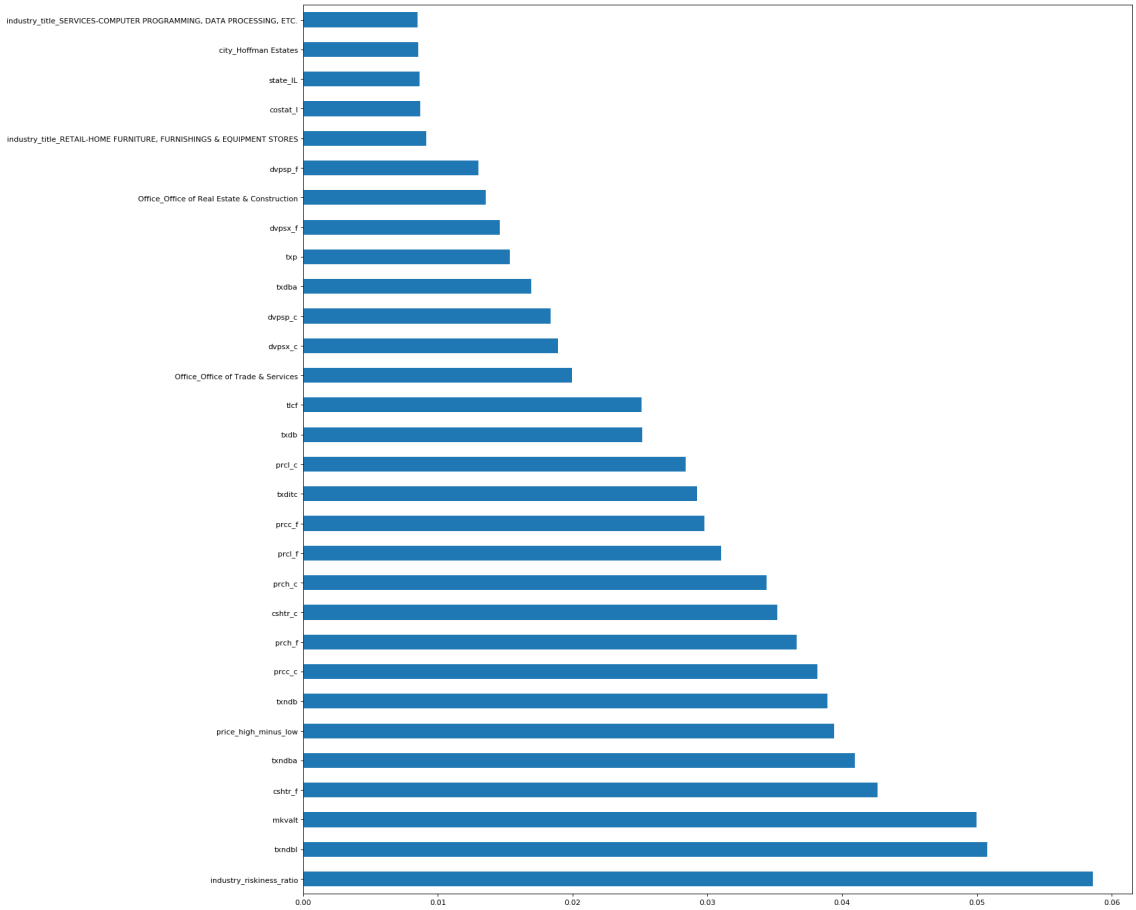

Figure 15: Feature Importance in Baseline Random Forest

precision suggests that the models are more judiciously picking good candidates for audits, rather than flagging a broad range of possibilities that exceed regulators' auditing capacity.

8 Discussion

8.1 Precision-Recall Tradeoff in Predictive Auditing

While this study is specific to cybersecurity, it speaks to a larger problem in law regarding government auditing to detect rare events. Governments frequently employ auditing as a tool to ensure that private actors are complying with regulations. In U.S. federal law, some common examples include Internal Revenue Service tax audits, Department of Labor fair labor standards audits, and Federal Emergency Management Agency disaster relief audits. These audits commonly target underlying activities that occur infrequently among legitimate activities. Most people adequately report and file their tax liabilities, most employers comply with fair labor standards, and most recipients of FEMA funds properly administer those funds. Indeed, a tiny percentage of each of these activities constitutes the sort of fraud or vulnerability that these audits are designed to uncover. Detecting these rare events is a problem because the government has limited resources to conduct audits. Given these constraints, governments may be concerned with ensuring that

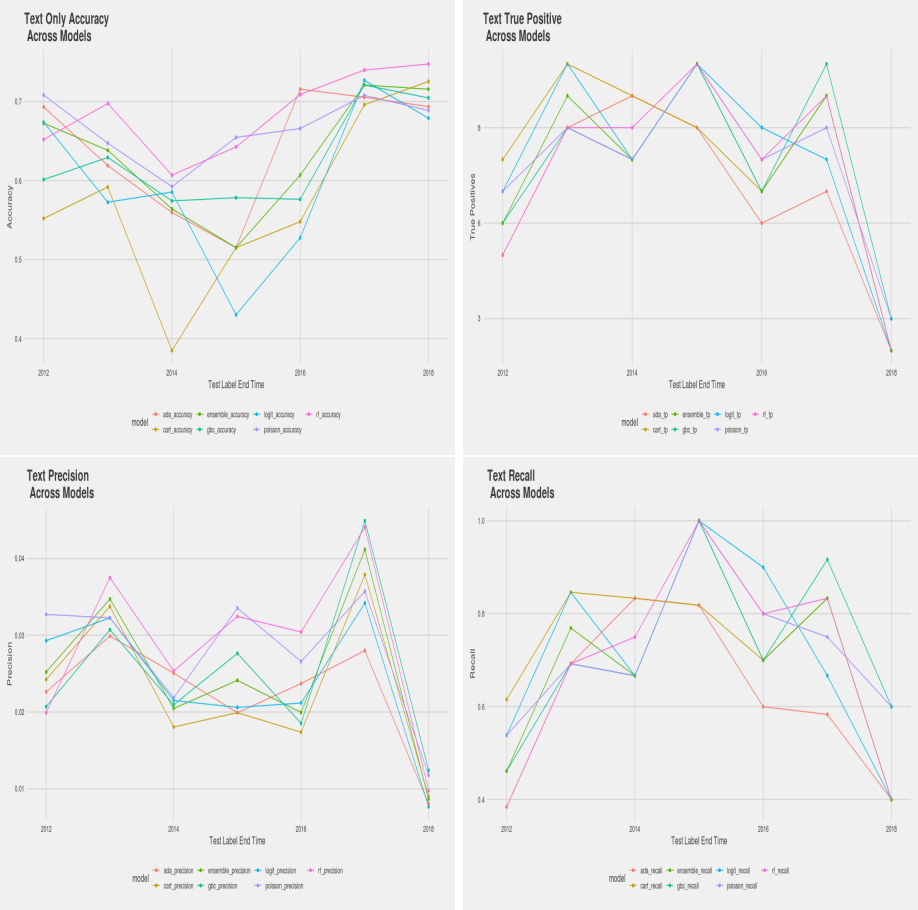

Figure 16: Text Only Results

auditing activity is directed towards undesirable activities.

In machine learning terms, this problem is best conceptualized as an imbalanced learning problem. Imbalanced learning refers to imbalance in the outcome variable of a dataset. In this cybersecurity context, the negative class ("no breach") swamps out the positive class ("breach"), as approximately only 2% of firms experience a breach each year. The core problem with imbalanced learning problems is that accuracy can be optimized simply by guessing the dominant class every time. However, when used to make to an actual decision, this type of model would not be useful. There are technical approaches to imbalanced learning problems, such as random over- and under-sampling as employed in this study. Thinking more broadly about how to map metrics to a policy context is also an important step though.

In policy contexts, precision and recall become relevant measures, but there is a tradeoff between them. One could achieve a perfect recall (finding all possible breaches) by assuming that every observation is a breach. However, the precision of this model would be quite poor, and if a government agency had the resources to audit every firm then an algorithmic approach would not be necessary. Similarly, a model could be very conservative and only make one guess about firms likely to be breached, and if that guess is correct, it could stop making predictions. While this approach would yield a perfect precision, it would miss many relevant cases, and again not be

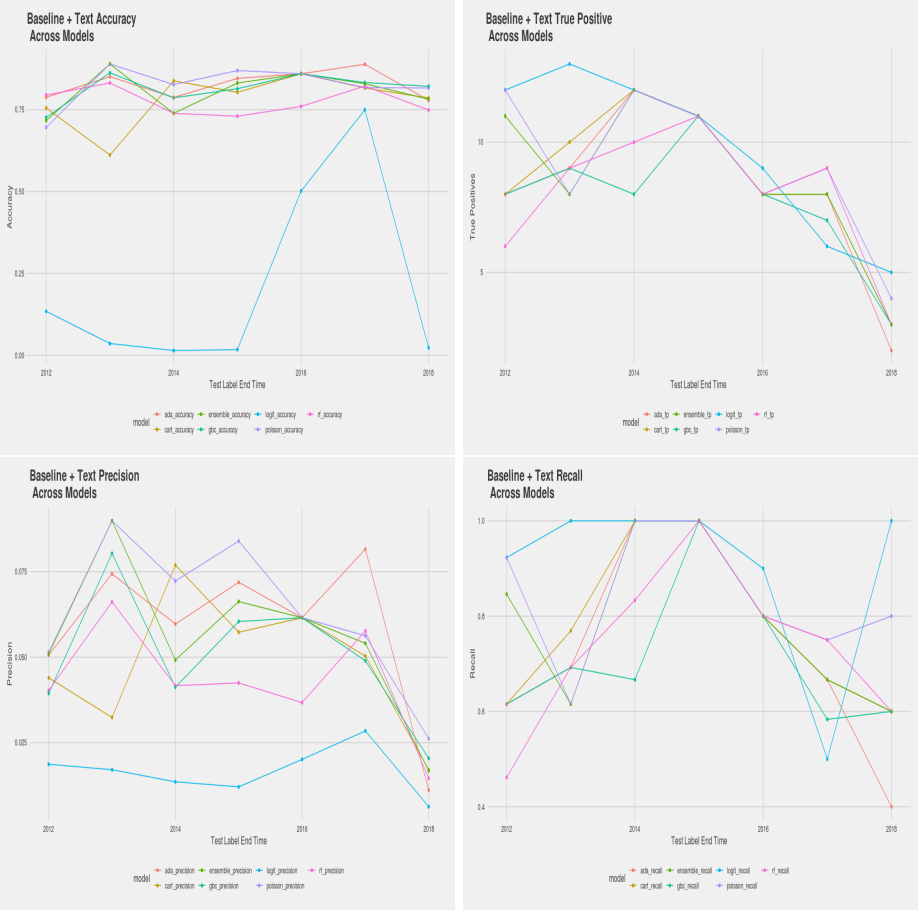

Figure 17: Baseline + Text Results

helpful for regulators who are trying to find the riskiest companies. This concept holds outside the cybersecurity context as well, as regulators frequently are implicitly optimizing the precision-recall tradeoff when targeting their auditing activities.

Framing the precision-recall tradeoff as part of a policy decision can help a decisionmaker determine the optimal amounts to trade off on each metric. In this cybersecurity context, a policymaker may prioritize maximizing true positives, maximizing recall, and minimizing false negatives, while tolerating weaker precision and a high number of false positives. These priorities are plausible because false negatives (failing to detect a breach) are more costly than false positives (auditing a firm that was not going to be breached). Similarly, recall (finding all potential breaches) may be more important than precision (the fraction of flagged firms that are actually breached). That being said, this tradeoff does not suggest optimizing these quantities by totally sacrificing precision for recall. Rather, contextualizing the tradeoff within the SEC's actual auditing program can help illuminate how policymakers should use these metrics.

To illustrate, assume that the SEC's auditing capacity is fixed at 75 audits per year. It will not conduct fewer than 75 audits even if doing so would be cheaper, nor does it have the resources to conduct more in a given year. Within these constraints, the SEC must optimize where to place these 75 audits to attempt to successfully detect companies that will be breached. Given that

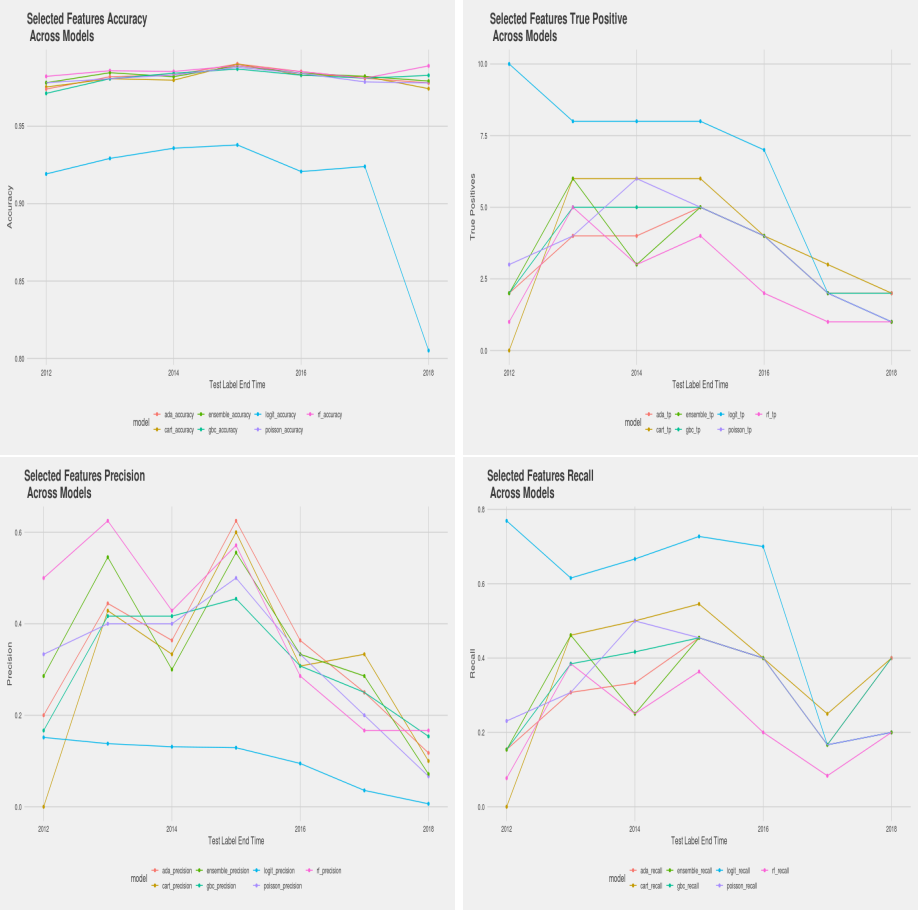

Figure 18: Selected Features Results

the number of audits is fixed, the precision is somewhat irrelevant. If a model flags 20 potential breaches, but only 5 were actually breached, the precision would be .25, but the audits of the 15 non-breached firms do not represent any marginal cost to the agency. Thus, the SEC may prioritize recall instead because it wants to make sure most of the riskiest firms do end up in the audit pool, as it will not have additional resources to audit those firms if they are not flagged. In situations where an agency wishes to conserve resources by reducing the number of audits, or the number of audits it makes is unbounded, prioritizing precision may be more sensible.

Figre 19 shows this tradeoff in the cybersecurity context. Using the predicted probabilities from the gradient boosting classifier model, it plots the precision-recall tradeoff. The "Random Audit" model guesses firms to pick for audits at random, and this model does quite poorly on precision. The GBC model on the other hand correctly flags several breached firms before guessing incorrectly. Importantly, while precision drops considerably once recall reaches about .5, auditors need not stop at that point. With 75 audits, the SEC could conduct audits up to a recall of about .71 before running out of resources. Thus, while auditors would tradeoff a considerable amount of precision with additional audits, doing so is not necessarily fatal to the enterprise as there are resources to spare in this case.

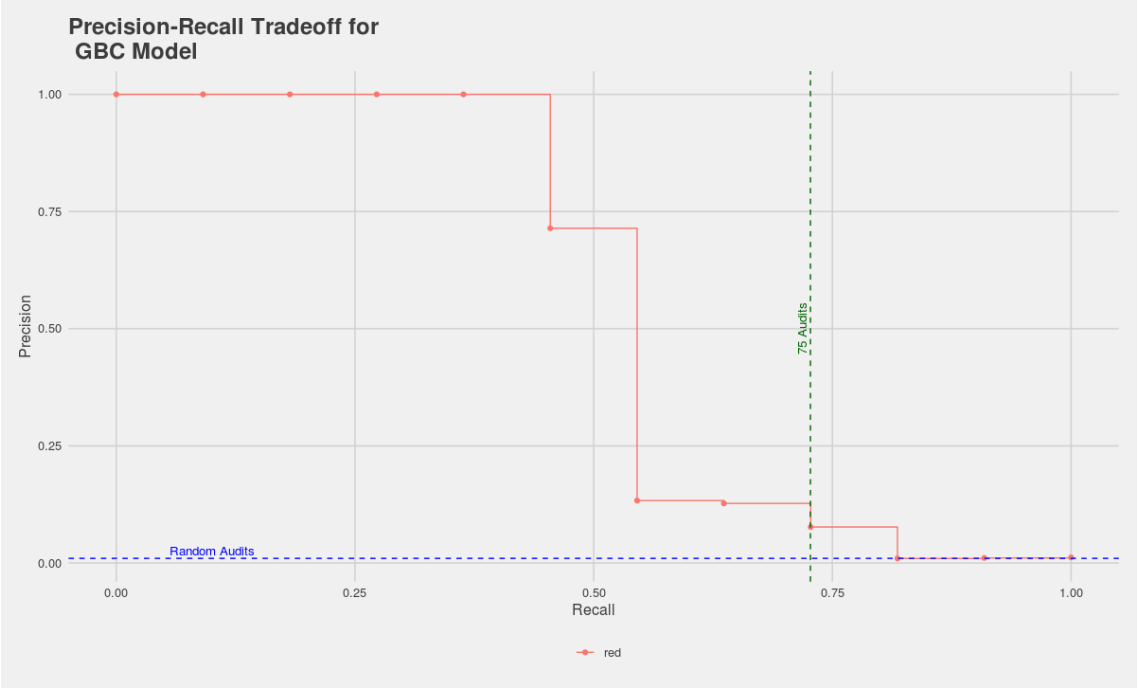

Figure 19: Precision-Recall Tradeoff

8.2 Simulation on Real-World Outcomes

I conclude with an illustration of how an algorithmic auditing approach improves upon a randomized approach. Figure 8 shows how many breached firms would be detected in advance for the 2015 fiscal year across 100,000 simulations. Occasionally, a regulator picking firms at random might find one breached firm, and rarely would find two. In most simulations, a random search would not yield any members of the positive class.

Figure 20 demonstrates the utility of an algorithmic approach over a randomized one. Using the assumption that SEC audits are totally effective at deterring a potential breach, it illustrates the potential reduction in breaches each year. Assuming 75 audits are conducted in each given year, we see an average reduction in breaches of about 18%. In the 2015 fiscal year, of the 24 breaches in the dataset, 5 are flagged in advance.

As seen in Figure 18, using the final models that select out unnecessary features, in most years the models achieve both recall and precision in the neighborhood of .4. While a poor precision score would generally be a problem in most machine learning applications, these results are actually quite promising when contextualized as a public policy problem. Although regulators would need to sift through several companies that are unlikely to be breached, the high recall suggests that they will eventually find companies that would have been breached and can act to bolster their cybersecurity practices. Most importantly, the algorithm eliminates a huge number of companies that it is confident will not be breached, thus saving regulators time and allowing them to focus their regulatory efforts on a smaller subset of companies. Table 5 illustrates this point with sample results from the ensemble algorithm's 2015/2016 predictions. A regulator could be furnished with a list that safely eliminates several companies from consideration, while allowing them to focus on

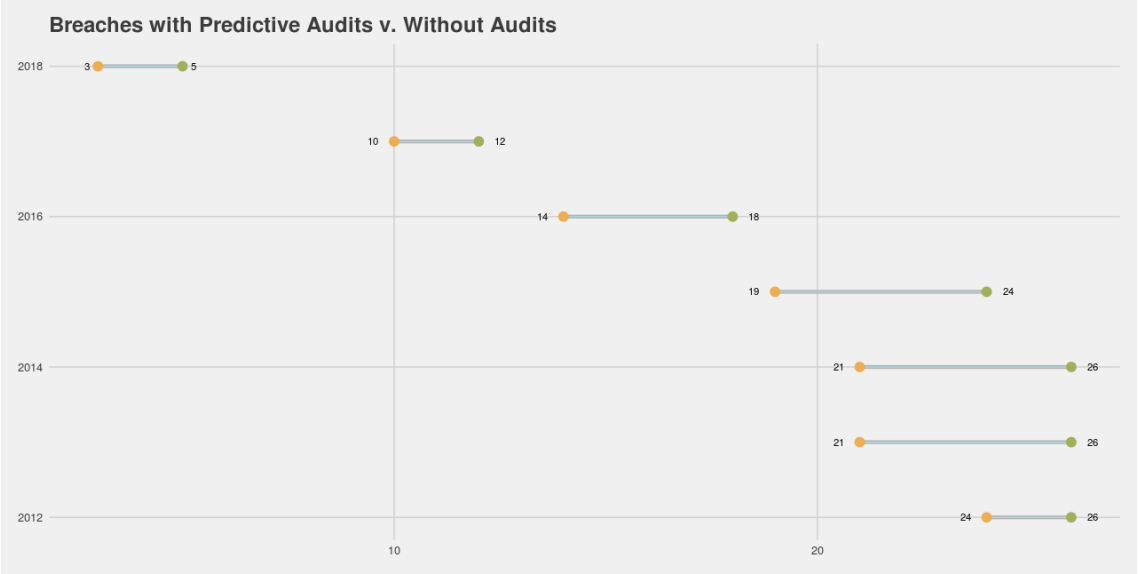

Figure 20: Breaches With and Without Predictive Audits

the likeliest breach targets.

The main normative takeaway for legal scholarship as a whole is that there is value to prediction. There is currently a live debate within law and law-adjacent literatures about the use of machine learning and prediction in legal contexts. Much of the attention thus far has understandably been placed on applications where decisions involve vulnerable populations and legally protected classiciations like race and gender. Thus, many of the examples focus on areas like employment, housing, and criminal law. This scholarly debate would be enriched by considering applications that do not implicate the same equity concerns. In this case, predicting the cyber-riskiness of corporations shares little similarity with the aforementioned examples on equity and fairness grounds. Instead, improving auditing efforts only improves efficiency, and is beneficial to regulators, corporations, and the public alike. Audits themselves are not costly for audited firms. While some firms may bear more of the costs of precaution, this allocation is sensible if they carry more of the risk. [Cite barocas/selbst, coglianese, lehr]

8.3 Simplifying Decisionmaking

Regulators may also choose to deploy simpler models that are more easily explained to outside stakeholders. Certain firm-level features are more predictive than others. For instance, a firm's industry's riskiness, location (New York, California, or Illinois), and stock volatility can be used to construct simple decision rules. These models can also incorporate flags for whether a firm's disclosure contains elements from the SEC interpretative guidance, and build a simple model to guide auditing decisions. For complex policy decisions, simplifying models can help with conveying the reasoning behind a legal decision. Simple models may sacrifice performance on certain metrics, but the added advantage of interpretability and ease of construction could be worthwhile. Jung et. al. detail this logic in depth. They argue for this "select-regress-and-round" approach. They

COMPANY NAME	filing date	breach	pred
hyatt hotels corp	2/18/2015	yes	yes
target corp	3/11/2016	yes	yes
chevron corp	2/25/2016	yes	yes
microsoft corp	7/31/2015	yes	yes
tennessee valley authority	11/20/2015	yes	yes
apple inc	10/28/2015	yes	yes
monster worldwide inc	2/11/2016	no	yes
iron mountain inc	2/26/2016	no	yes
quest diagnostics inc	2/26/2016	no	yes
commercial metals co	10/30/2015	no	no
medallion financial corp	3/11/2015	no	no
marriott international inc	2/19/2015	no	no

Table 5: Sample Results for Predicting 2016 breaches. The "breach" column indicates firms that were actually breached in 2016, "pred" indicates firms that were predicted to be breached in 2016. Blue indicates a "breach" value and red indicates a "no breach" value.

advocate for a pipeline where the analyst builds a complex model that serves as a benchmark, and then create simple rules to test against both this benchmark and human decisions. They highlight the use of simple rules in judges making bail decisions, and note that simple rules both outperform human judges and come close to complex models like random forests (Jung et al., 2017).

In the cybersecurity context, we can see the value of this framework by using a decision tree and benchmarking it against the ensemble model. Figure 21 illustrates a classification tree built with these features alone on the same 2015-2016 period used above. The basic logic of the tree makes splits based on market value and industry riskiness to make predictions about whether a particular observation is a "breach" or "no breach." Although the model does slightly worse on recall than more complex models, it still does relatively well and is much simpler to visualize. In lieu of the more complex models used earlier, the SEC could choose to use a simple classification tree with some manually engineered features to achieve comparable results. Importantly, this simple model still avoids the accuracy trap of flagging everything as "no breach," and the recall trap of flagging everything as "breach," as seen in Table 6.

One way to approach this problem would be to start with the more complex models described above, and then map their complex decision rules to simple ones for deployment in practice. The exploratory analysis and complex modeling helped surface insights into which features were genuinely informative, the types of mistakes that different modeling choices would lead to, and the best possible performance of a model in this context. From these complex models, it is possible for a regulator to narrow down the features to prioritize, and focus on creating a decisionmaking pipeline that utilizes that simpler information (as I do here with a classification tree). This simpler

model can be used to explain the process and justification for important legal decisions, even if more complex models were fit first. Critically, these models should not be static. Observing real-world outcomes, adjusting regulations, and retraining models should be a dynamic process that informs human decisionmaking in policy contexts, not replaces it.

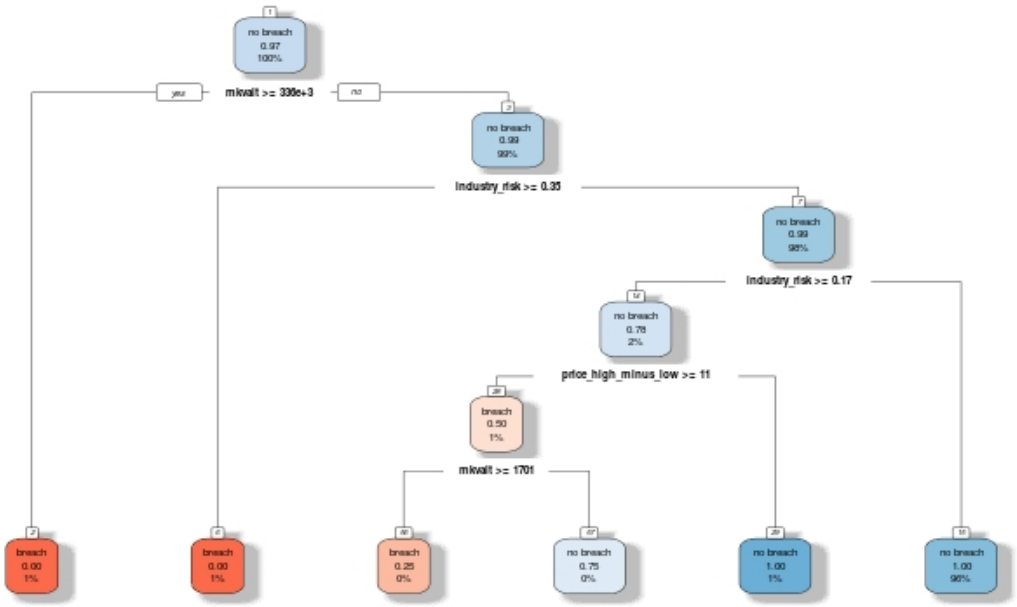

Figure 21: Classification Tree

	observed breach	observed no breach
predicted breach	9	13
predicted no breach	9	872

Table 6: CART Confusion Matrix

9 Future Work

There are several areas of improvement for future work to iterate upon these results. These results are drawn from matching outcome data from public databases to publicly traded companies. However, this construction is not complete. For instance, the SEC flagged 87 breaches in 2017, compared to the approximately 20 breaches I found for the same year by manually cross-checking publicly reported breaches to companies in the dataset. Resolving these inconsistencies would help bring the models closer to the ground truth, and likely help the class imbalance problems as well.

A qualitative component that includes discussions with SEC auditors, firm managers, and in-house cybersecurity personnel would also be helpful. Many of the assumptions about how managers word their cybersecurity disclosures and report their cyberrisks are based on theoretical reasoning

and second-hand sources. Gaining more insight into how disclosures are actually crafted, and how companies think about their own cybersecurity postures, would be tremendously helpful in building better models. Moreover, speaking to regulators and learning what their priorities are would help determine which metrics to prioritize, and how to target audits. In particular, gaining more insight into the exact mechanism underlying the current choice of firms to audit would help establish a realistic baseline beyond random audits.

Finally, a field experiment that validates the modeling would be invaluable. While the temporal cross-validation provides some evidence of how the model would have worked historically, this is not a guarantee of performance in the future. Randomizing firms flagged by the model and observing differences in breach rates would validate the model's predictions. Creating an interplay between training new models and real-world testing will ensure that the models stays up-to-date and usable. Most importantly, targeting interventions at the firms most likely to benefit from them creates an opportunity to assess the causal effect of the audits themselves, and reevaluate the SEC guidelines and audits in light of quantitative evidence.

10 Conclusion

Like with many policy areas, privacy and cybercrime scholarship has traditionally focused on the theoretical underpinnings of causation. This study looks to expand the traditional scholarship by reframing cybersecurity as a prediction policy problem. Predicting incidents before they occur gives policymakers and organizations many more opportunities to prevent the privacy harms that stem from massive data losses. Prevention would be more effective than restitution, and tools that can aid in this goal would reshape the current discourse around data protection law that focuses mainly on harms.

If successful, this study could also bolster current efforts to incorporate artificial intelligence and data science into regulatory efforts. Mandatory disclosure is a commonly used and powerful legal mechanism for ensuring better institutional behavior. Scholars and policymakers have extolled the virtues of disclosure for decades. New computational tools potentially allow us to harness not only the fact that a disclosure is made, but the actual content of a disclosure. Incorporating data science into the framework of disclosure law could spur a flurry of innovative scholarship. Tools that make sense of the massive amount of text generated by mandatory disclosure can improve regulatory efforts, increase consumer information, and promote healthier corporate behavior.

References

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*.

Bauguess, S. W. (2017, June). The Role of Big Data, Machine Learning, and AI In Assessing Risks: A Regulatory Perspective.

Brandeis, L. D. (1914). *Other People's Money and How the Bankers Use It*. Frederick A. Stokes.

Cowley, S. (2017, October). 2.5 Million More People Potentially Exposed in Equifax Breach. *The New York Times*.

for Data Science, C. and P. P. at the University of Chicago (0). Temporal Cross-Validation.

Jung, J., C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein (2017, April). Simple Rules for Complex Decisions. *arXiv:1702.04690v3*.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction Policy Problems. *American Economic Review*.

Kogan, S., D. Levin, B. R. Routledge, and J. S. Sagi (2009, June). Predicting Risk from Financial Reports with Regression. *Human Language Technologies: The 2009 Conference of the North American Chapter of the ACL*.

Lapowsky, I. (2018, April). Facebook Exposed 87 Million Users to Cambridge Analytica. *WIRED*.

Leuz, C. and P. Wysocki (2016, February). The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestion for Future Research. *Journal of Accounting Research*.

Liu, Y., A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, and M. Liu (2015). Cloudy With a Chance of Breach: Forecasting Cyber Security Incidents. *Proceedings of the 24th USENIX Security Symposium*.

Mayer, J. (2016). Cybercrime Litigation. *University of Pennsylvania Law Review*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013, September). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301:3781v3*.

Mitts, J. (2014). Predictive Regulation. *SSRN Working Paper*.

Mitts, J. and E. Talley (2018). Informed Trading and Cybersecurity Breaches. *Harvard Business Review*.