

Risk Management

Entity-Specific Cyber Risk Assessment Using Insurtech-Empowered Risk Factors

Jiayi Guo^{1a}, Linfeng Zhang^{2b}, Zhiyu Quan^{1c}

¹ Department of Actuarial and Risk Management Sciences, University of Illinois Urbana-Champaign, ² Department of Mathematics, The Ohio State University

Keywords: Risk assessment, Cyber risk, Machine learning, Insurtech

Variance

Vol. 19, 2026

The lack of high-quality public cyber incident data limits empirical research and predictive modeling for cyber risk assessment. This challenge persists because companies are reluctant to disclose incidents that could damage their reputation or investor confidence. From an actuarial perspective, potential resolutions include these: the enhancement of existing cyber incident datasets and the implementation of advanced modeling techniques to optimize the use of the available data. A review of existing data-driven methods highlights a significant lack of entity-specific organizational features in publicly available datasets. To address that gap, we propose a novel insurtech framework that enriches cyber incident data with entity-specific organizational features. We develop various machine learning (ML) models: a multilabel classification model to understand the occurrence of cyber incident types (e.g., privacy violation, data breach, fraud and extortion, IT error, and others) and a multi-output regression model to estimate their annual frequencies. While classifier and regressor chains are also implemented to explore dependencies among cyber incident types, no significant correlations are observed in our datasets. We also apply multiple interpretable ML techniques to identify and cross-validate potential risk factors developed by insurtech across ML models. We find that compared with conventional risk factors, insurtech-empowered features enhance occurrence and frequency estimation robustness. The framework generates transparent, entity-specific cyber risk profiles, supporting customized underwriting and proactive cyber risk mitigation. It provides insurers and organizations with data-driven insights to support decision-making and compliance planning.

a Jiayi Guo is a PhD student in the department of actuarial and risk management sciences of the University of Illinois at Urbana-Champaign. Before joining the doctoral program, she earned her master's degree in predictive analytics and risk management from University of Illinois at Urbana-Champaign. Her research focuses on risk modeling using machine learning and data-driven methodologies, with a particular emphasis on climate risk within actuarial science. She currently serves as a graduate supervisor for the climate data innovation and consolidation project at the Illinois Risk Lab.

b Linfeng Zhang is an assistant professor in the department of mathematics at The Ohio State University. He earned his PhD in mathematics, with a concentration in actuarial science and risk analytics, from the University of Illinois at Urbana-Champaign. He served as a research assistant at the Critical Infrastructure Resilience Institute from 2016 to 2019, focusing on cyber risk and cyber insurance research. His research on cyber and privacy risks has been funded by SOA, CAS, and Cisco. Linfeng has published research in leading actuarial science and risk management journals, including JRI, ASTIN Bulletin, and IME.

c Zhiyu Quan is an assistant professor in the department of actuarial and risk management sciences of the University of Illinois Urbana-Champaign, a Brad and Karen Smith Professorial Scholar of the College of Liberal Arts and Sciences, and a finance and insurance sector lead for Discovery Partners Institute. He holds a PhD in actuarial science from the University of Connecticut. Before joining Illinois, he worked in a cutting-edge insurtech company as an R&D data scientist developing data-driven solutions for major insurance companies. He has a broad spectrum of research interests in data science applications in insurance and applies his actuarial expertise to build predictive models. His research is driven by real-life data and is inspired by collaborations with insurtech and insurance companies. He also is a director of the Illinois Risk Lab and has received the Arnold O. Beckman Research Award and awards from the Society of Actuaries Research Institute and Casualty Actuarial Society. Address for correspondence: zquan@illinois.edu.

1. INTRODUCTION

Cyber incidents, such as data breaches and ransomware attacks, pose significant threats to modern businesses. The consequences of operational cyber risk can compromise the confidentiality, availability, or integrity of information or information systems and ultimately result in financial loss (see Cebula et al. (2014)). To mitigate such impacts, many organizations have turned to cyber insurance, which has become an integral component of enterprise risk management. From the insurance industry's perspective, cyber risk assessment is a crucial component in underwriting practices, and it typically relies on historical experiences with incidents that caused material losses. However, as identified in Biener et al. (2015), the scarcity of data on historical cyber incidents is a major barrier that hinders the development of the cyber insurance market because it prevents pricing models from being effectively evaluated. More recent studies, such as that of Zängerle and Schiereck (2023), suggest that although marginal improvements have been made over the years, the shortage of data remains an issue. As well as insurers, organizations in need of a better understanding of their own cyber risks also have difficulties in learning from the past to support the assessment. Moreover, the lack of quality data in the public domain limits the ability of researchers to generate insights from empirical studies and build high-performing predictive models.

1.1. DATA SCARCITY ISSUE

The scarcity of data is unlikely to be addressed in the short term. To explain why, Romanosky (2016) describes the typical data-generating process used to create existing public or proprietary cyber incident databases. That process consists of a sequence of three actions: the detection, reporting, and recording of a cyber event after the event occurs. Cyber incident detection is a technically challenging task because the consequences of many incidents, such as data breaches, are not immediately tangible. A survey by IBM (2024) finds that the average time to detect a data breach is more than 200 days. The reporting of cyber incidents is largely a regulatory problem because companies lack the incentive to report incidents, as such information may very likely deliver a negative signal to customers and investors. To highlight regulators' role in improving cyber incident data, Kesan and Zhang (2021) note that the number of recorded data breaches in the US grew substantially after states started to enact data breach notification laws. Incident detection and reporting determine the number of disclosed cyber events that are available to be recorded, and as long as the technical and legal challenges noted persist, an ample amount of cyber incident data will be unavailable. Thus, under the circumstances, the questions that we may be able to address from an actuarial perspective include how we can enrich the existing cyber incident data and how we can apply advanced modeling techniques to the existing data effectively for cyber risk assessment. In the following, we offer an overview of the proposed data-driven approaches in this regard in the literature.

1.2. DATA-DRIVEN APPROACHES TO CYBER RISK MODELING

Several cyber incident datasets are available in the public domain, and many studies leverage them for cyber risk modeling with a focus on the frequency and severity of various types of incidents and the dependence among them. For example, Wheatley et al. (2021) use information on data breach events from the Privacy Rights Clearinghouse (PRC) and the Open Security Foundation to statistically model the frequency and severity of major breaches with different causes. Also based on the PRC data breach dataset, Eling and Jung (2018) use copulas to model the dependence among cyber incidents of different kinds and across various industries. Xu et al. (2018) use PRC's dataset to estimate time series models for the inter-arrival time and size of breaches and to model the dependence between them using copulas. Kesan and Zhang (2020) use a proprietary cyber incident dataset compiled by Advisen to study the key factors that drive the litigation following security incidents and the outcomes of lawsuits. Kesan and Zhang (2021) combine Advisen data with governmental information technology (IT) budget information to model the impact of IT spending on cyber risk. Palsen et al. (2020) use Advisen data and random forest classifiers and regressors to predict the legal and financial outcomes of cyber incidents based on characteristics of incidents and affected organizations. Romanosky (2016) builds linear models on Advisen data to study the factors influencing the financial impacts of cyber incidents on organizations. Eling and Wirfs (2019) extract cyber-related incidents from an operational risk database, SAS OpRisk Global Data, and model the distribution of cyber losses in relation to incident and company characteristics.

It is worth noting that there are relatively few studies on entity-specific risk modeling because of the general lack of features associated with individual affected organizations in most of the available datasets, especially the ones in the public domain. For example, Wheatley et al. (2021) model the frequency and severity of data breaches at the industry level. In practice, according to Romanosky et al. (2019), cyber insurance policies are often priced based on very few objective factors, including revenue and industry classification, in combination with expert judgment. From an underwriting perspective, this level of granularity may not be sufficient due to the significant variability of riskiness associated with organizations of similar sizes and in the same industry but with otherwise different risk characteristics.

In addition, whereas dependence among different incident types is also often modeled at the industry level (see, for example, Eling and Jung (2018)), dependence at the entity level remains unknown. That is, does being susceptible to one type of incident suggest that the organization is prone to other types of incidents? This question relates to designing multiple-peril cyber insurance as discussed in Chong et al. (2025). It is common practice for cyber insurance policies to cover multiple types of incidents, such as data breaches and ransomware attacks, and therefore a better understanding of entity-specific dependence is crucial.

1.3. CYBER RISK MODELING WITH AN ENRICHED DATASET

In this study, we define a cyber incident as any event that compromises the confidentiality, integrity, or availability of entity-specific information systems, and we focus on modeling such incidents at the entity level to enrich existing datasets. Based on existing data and literature, we propose that cyber incident data, despite its scarcity in terms of number of records, can be enriched by entity-specific information of victim organizations, enabling a more thorough and detailed analysis of how the risk factors of an organization affect its cyber risk.

We make our best effort to collect cyber incident data from the public domain, which we then augment with additional data from a proprietary insurtech source that offers a rich set of features associated with individual organizations. This collective dataset enables us to build ML models for the assessment of entity-specific cyber risk and understand the occurrence and frequency of cyber incidents that an organization could possibly experience. The models are designed to answer the following questions:

1. Compared with the conventional insurance risk factors used in practice, such as industry and revenue, do additional entity-specific organizational features offer an improved incident evaluation performance?
2. Beyond capturing the occurrences of various incidents (i.e., whether an incident of a particular type would occur), can their frequencies (i.e., the number of incidents of a certain type) be evaluated reasonably well with the currently available data?
3. At the entity level, when the firm-specific characteristics are given, is there any conditional dependence among the occurrences and frequencies of various kinds of incidents?

Our findings suggest that, besides revenue and industry, alternative entity-specific organizational insurtech-empowered features, such as the customer reviews of a business, indeed improve the performance of cyber risk models. The models also reveal that, with a rich set of entity-specific organizational features, the occurrence of various types of cyber incidents can be captured with reasonable accuracy, though evaluating their frequency remains a challenge. Finally, we find no evidence of entity-level dependence among the occurrences or frequencies of different types of cyber risks. We base this observation on the result that the classification and regression models with predictions on one type of incident being conditioned on another fail to prevail over the ones that treat individual incident types as independent. Note that this result does not challenge the intuition that firms with certain traits, such as strong cybersecurity, may experience lower occurrence and frequency across all types of cyber incidents. Rather, it suggests that the occurrences or frequencies of different types of incidents are likely to be conditionally independent, given the firm's characteristics.

The contributions of our work are as follows. The study offers a comprehensive review of cyber incident datasets in the public domain to facilitate future research in the

field. That incident data is enriched by entity-specific or organizational insurtech-empowered features, thus enabling risk and dependence modeling for individual organizations. The study further examines the effectiveness of entity-specific organizational features in capturing cyber incident occurrence and frequency and reveals a set of potential rating factors for cyber insurance. Lastly, the study suggests that dependence among different cyber incident types may not be present at the entity level.

The remainder of the paper is structured as follows. In Section 2, we introduce several publicly available cyber incident datasets, the insurtech data that offers additional entity-specific risk characteristics, and the comprehensive dataset that compiles data from those sources and that we use for the study. Section 3 details the modeling methodology we adopt to answer the three aforementioned questions, and Section 4 offers a summary of the performance of various models and the comparison among them. The feature importance results derived from the proposed models are described in Section 5. In Section 6, we discuss the results and implications of this study and conclude.

2. DATA

To assess cyber risk at the individual-entity level, we leverage data and analytical tools made possible by recent advances in insurtech. Our approach incorporates entity-specific organizational digital footprints, social media activity, and other emerging data sources increasingly accessible through insurtech. This enables the development of more granular and dynamic risk assessments. We begin by identifying key outcome data related to past cyber incidents, such as incident type, followed by forward-looking risk factors that capture an entity's exposure to cyber threats, such as metrics associated with the entity's public image.

2.1. CYBER INCIDENT DATA SOURCES

We offer a summary of the cyber incident datasets that exist in the public domain, including the aforementioned PRC data breach database and some others that are less commonly seen in academic studies. Despite variations in details captured by different data sources, all of the datasets we consider in this study contain at least some description of the nature of the incident, e.g., whether it is a data breach or a ransomware attack, the year the incident took place, and the entity affected by the incident.

PRC's database is a collection of publicly reported data breach incidents in the US (see Privacy Rights Clearinghouse (2025)). It gathers data security events disclosed by entities that are subject to the HIPAA Breach Notification Rule and/or state data breach notification laws. PRC pools the reported events from the US Department of Health and Human Services (HHS) and state law enforcement agencies, extracts additional information from the reported events, such as organization and breach type, and compiles the data from various sources into a standardized tabular format. The archive of incidents between 2005 and 2019, which contains more than 9,000 incidents, is freely avail-

able.¹ Records of more recent incidents have to be purchased. Because PRC collects information from public sources, it is possible to manually and directly retrieve the recently disclosed incidents from HHS and states that publish such data, such as California and Oregon, but collecting data from those sources individually requires significant effort.

The VERIS Community Database (VCDB) is a community-maintained public database of security incidents that follows the Vocabulary for Event Recording and Incident Sharing (VERIS) data model (see Verizon RISK Team (2025)). The VERIS schema specifies the information to be recorded for each cybersecurity incident in standard formats, including victim demographics, incident description, responses, and outcomes. Actively maintained at the time of this writing, the database contains more than 10,000 records contributed by community members. However, because the records are voluntarily reported or collected, some potentially come from unverified sources and may have accuracy issues. The VERIS framework includes a confidence rating for the accuracy of each record. About 3,000 of the 10,000 records have a confidence rating of “medium” or “high.”

The Center for International and Security Studies at Maryland (CISSM) maintains a Cyber Events Database (the CISSM database) (see Harry and Gallagher (2018)). A research team from the University of Maryland’s School of Public Policy scrapes websites that publish information on cybersecurity incidents, manually reviews the gathered data, and documents the incidents in a tabular format. The data collection is performed monthly, and at the time of this writing, the number of recorded incidents in this database is over 14,000.

The aforementioned three datasets are among the largest cyber incident datasets in the public domain. In addition, there are some smaller but notable ones. The Cybersecurity in Application, Research and Education Lab at Temple University maintains a Critical Infrastructure Ransomware Attacks dataset that collects data on publicly disclosed or reported ransomware attacks. It currently has more than 2,000 records (see Rege, Bleiman, et al. (2023)). Researchers at the University of Queensland created a dataset of data breaches and ransomware attacks spanning a period of 15 years, starting in 2004. The dataset has over 1,000 recorded incidents (see Ko et al. (2020)). Have I Been Pwned, a dataset that keeps track of websites that have been breached and the amount and type of leaked data, currently includes data from more than 800 incidents (see Hunt et al. (2025)).

In addition to the structured datasets discussed above, another valuable source of information on historical security incidents is Doe (2025), which regularly publishes brief articles covering recent cyber incidents and other cybersecurity-related news, such as regulatory updates. Since 2008, the platform has published more than 35,000 articles, mak-

ing it a rich repository of unstructured text data. Due to its unstructured nature, however, extracting meaningful insights requires advanced natural language processing (NLP) techniques. Inspired by our prior work in cyber literature collection and NLP, we apply a text analysis pipeline to extract structured information from these articles (see Zhang et al. (2025)). Specifically, we utilize DeepSeek-R1 7B, a pretrained large language model (see DeepSeek-AI et al. (2025)), to identify and extract several key attributes from each article, including (1) whether the article describes a cyber incident, (2) the name of the affected entity, (3) the date of the incident if it is a recent incident, (4) whether the article is a follow-up report on an incident in the past, and (5) the type of the incident. This approach produced more than 11,000 identified incidents with known incident dates, incident types, and affected organizations.

In this study, for each cyber incident dataset, we keep only the names of affected organizations, incident dates, and incident types. The originally documented incident types vary from dataset to dataset. For example, the CISSM database categorizes incidents into four groups: disruptive, exploitive, mixed, and undetermined. Disruptive and exploitive incidents are further categorized into 10 subgroups. In comparison, each incident in the VCDB is associated with one or more of six actions that led to the event, such as hacking and error-related causes, and one or more affected assets from a comprehensive list, such as database servers and network infrastructure. To unify these different approaches to incident classification, we follow the four categories proposed in Kesan and Zhang (2021)—i.e., data breach, privacy violation, extortion/fraud, and IT error—and map the existing set of categories used by each cyber incident dataset to them. Incidents that cannot be classified are labeled as *other*.

Note that these data sources have overlapping records. For example, the mass data breach experienced by Equifax in 2017 appears in multiple aforementioned datasets. To create a single combined dataset for this study, we check for duplications under the assumption that an organization does not experience more than one incident of the same type on the same day, and duplicated records are removed.

2.2. INSURTECH DATA

Traditional cyber insurance underwriting largely relies on conventional risk factors such as industry classification, business size, annual revenue, and responses to IT security surveys, as summarized by Tsohou et al. (2023), Nurse et al. (2020), and Romanosky et al. (2019). While such inputs offer a general overview, they often fall short of capturing the true nature and severity of an organization’s cyber risk. In many cases, the information is either too generic or misaligned with the way attackers actually assess targets. For instance, survey-based questions such as “How much does your company spend annually on IT security?” may not

¹ <https://privacyrights.org/sites/default/files/2020-01/PRC Data Breach Chronology - 1.13.20.csv>.

yield accurate or meaningful answers. IT budgets are often approved at a broad entity-specific organizational level, making it difficult for respondents to isolate security-specific expenditures. Moreover, some policyholders may give incomplete or intentionally misleading responses, whether due to uncertainty, oversight, or misaligned incentives. As a result, survey-derived data may lack the precision and reliability needed for effective underwriting. As Romanosky et al. (2019) point out, the connection between potential cyber losses and the collected security information is unlikely to be quantitatively determined. Also, as Tsohou et al. (2023) suggest, the lack of an objective way to assess cyber risk contributes to the difficulties in assessing such risk, and the cyber insurance industry is in need of a risk assessment standard beyond the customized questionnaires that are commonly seen in current practice.

To address these shortcomings, it is critical to incorporate more objective and externally observable indicators of cyber risk. Insurtech platforms now enable access to digital signals such as an organization's online footprint, social media activity, exposed credentials, domain registration history, unpatched software vulnerabilities, and public-facing infrastructure configurations. These external data points provide a more accurate and real-time reflection of an entity's risk posture from the perspective of potential attackers.

For this study, we acquire a dataset of over 500 entity-specific organizational features from an insurtech platform, Carpe Data.² These features encompass a comprehensive range of business indicators, including geographic information, enterprise structure, operating status, customer review metrics, publicly disclosed business risk characteristics, classification segment tags, firmographics (e.g., size, revenue, employees), licensing categories, open hours statistics, and minority shareholding information. A detailed description of the insurtech dataset can be found in Quan et al. (2024). To maintain conciseness, we omit a comprehensive overview here and instead introduce relevant data components as they pertain to our discussion. Briefly, we provide a category-level summary of the 527 insurtech features we use in this study in Appendix D, [Table 4](#).

Notably, this metadata is not derived from internal cyber incident records but from publicly accessible digital and operational attributes of the enterprise. Many of these features have not been examined in the cyber risk literature, and we shall later show that some of them—such as several key metrics associated with a firm's customer reviews—significantly affect the cyber risk of a firm.

2.3. ASSEMBLED DATA

The final datasets used for both multilabel classification and multi-output regression are constructed through a comprehensive data integration and processing pipeline,

incorporating two primary sources: cyber incident data and insurtech data.

The cyber incident dataset is first compiled using the data sources described in Section 2.1 and additional incident records from a proprietary source. It includes three features: *COMPANY_ID* (the unique company identifier), *CASE_TYPE_LG* (the type of cyber incident), and *ACCIDENT_YEAR* (the year the incident occurred, ranging from 1903 to 2018). Following the removal of records with missing identifiers or years, and after applying a multistage deduplication process, the dataset is reduced to 66,215 valid entries.

To prepare this data for modeling, the categorical incident type variable *CASE_TYPE_LG* is converted into a multilabel format using one-hot encoding, resulting in five binary indicators corresponding to distinct types of cyber incidents: (1) privacy violation, (2) data breach, (3) extortion/fraud, (4) IT error, and (5) other. Each row in the cyber incident dataset corresponds to a single reported incident. To study how insurtech-empowered features relate to the occurrence and frequency of cyber incidents, we restructure the data at the firm-year level, where each observation represents a single firm in a single observed calendar year and consolidates all incidents experienced by that firm during that calendar year. Specifically, to obtain this data structure, the five binary indicators are aggregated accordingly at the firm-year level, collapsing multiple incident-level entries into a single firm-year summary. For instance, if a firm experienced two privacy violations and one IT error in 2019, its aggregated frequency vector becomes

$$\begin{aligned} \textit{Privacy Violation} &= 2, \\ \textit{Data Breach} &= 0, \\ \textit{Extortion / Fraud} &= 0, \\ \textit{IT Error} &= 1, \\ \textit{Other} &= 0. \end{aligned}$$

This transformation ensures that the output variables reflect the frequency with which each incident type occurred within the year. After applying this transformation across all firms and calendar years, the dataset contains a total of 53,243 unique firm-year observations, each representing the full distribution of cyber incident types experienced by a company in a given calendar year. The aggregated frequency counts also allow us to construct the binary occurrence indicators by mapping any positive count to 1.

In parallel, the insurtech partner contributes a rich dataset of 528 potential risk factors covering 31,782 companies. To integrate the cyber incident records with the insurtech dataset, we align the two datasets at the firm-year level by merging observations through their shared company identifier: *COMPANY_ID*. However, due to limitations in historical data coverage, specifically the lack of company identifiers for incidents before the 2000s, the information for incidents occurring in the early 20th century is excluded during alignment. After retaining only firm-year

² <https://carpe.io/>.

pairs that appear in both datasets and removing identifier-related variables, the final modeling dataset contains 39,636 firm-year observations with 533 columns. Among these, 528 are features: 527 from the insurtech dataset and one from the cyber incident dataset (*ACCIDENT_YEAR*). The remaining five columns constitute the multilabel output variables.

Formally, for the multilabel classification problem, we define the dataset as

$$\mathcal{D}_{\text{cls}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m,$$

where each input $\mathbf{x}_i \in \mathbb{R}^d$ represents a d -dimensional feature vector, and m denotes the total number of observations in the dataset. For insurtech-empowered features, the vector can be written as

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id}), \quad i \in \{1, 2, \dots, m\},$$

where $d = 528$ and $m = 39,636$.

For the output variable space $\mathcal{Y} = \{0, 1\}^q$, each observation has a q -dimensional output variable vector $\mathbf{y}_i$ from the label set, $\mathcal{L} = \{L_1, L_2, \dots, L_q\}$, with each L_j representing a unique label, the index $j \in \{1, 2, \dots, q\}$ preserving the order of labels across all observations. In our case, the label set $\mathcal{L} = \{\text{Privacy Violation, Data Breach, Extortion/Fraud, IT Error, Other}\}$, where $q = 5$. The output variable vector $\mathbf{y}_i$ serves as a binary indicator of the presence (1) or absence (0) of each L_j for i -th observation and can be written as

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iq}), \quad i \in \{1, 2, \dots, m\},$$

where $q = 5$ and $m = 39,636$.

Specifically, the j -th element of $\mathbf{y}_i$ can be written as

$$y_{ij} = \begin{cases} 1, & \text{if the label } L_j \text{ presents in } i\text{-th observation} \\ 0, & \text{otherwise} \end{cases}, \quad j \in \{1, 2, \dots, q\}.$$

Hence, we can have multiple incident types present in a single firm-year observation.

Similarly, for the multi-output regression problem, we define the dataset as

$$\mathcal{D}_{\text{reg}} = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^m,$$

where each output variable vector $\mathbf{z}_i \in \mathbb{R}^q$ contains the count of cyber incidents in each of the q categories for firm i during a specific year, where the observation size m remains consistent with the classification setting. The output dimensions are denoted by $\mathcal{O} = \{O_1, O_2, \dots, O_q\}$, corresponding to the same categories as in the classification task. Thus, the output variable vector $\mathbf{z}_i$ represents the number of times cyber incident type O_j occurred for the i -th observation. Specifically, the j -th element of $\mathbf{z}_i$ can be written as

$$z_{ij} \in \mathbb{R}, \quad i \in \{1, 2, \dots, m\}, \quad j \in \{1, 2, \dots, q\}.$$

For both modeling tasks, the dataset is randomly partitioned into training and test subsets using a fixed split ratio of 80% and 20%, respectively, ensuring consistent evaluation across methods.

3. METHODOLOGIES

3.1. MULTILABEL CLASSIFICATION

Multilabel classification is a supervised learning framework in which each observation can be simultaneously associ-

ated with multiple labels, rather than being limited to a single mutually exclusive category. This approach is particularly relevant in actuarial applications where complex relationships exist; for example, a single policyholder may hold multiple types of cyber insurance coverage with a company. In such cases, the traditional assumption of one label per observation does not hold. The multilabel setting introduces additional modeling challenges, including the need to account for dependencies among labels and to address issues related to label sparsity and class imbalance, both of which are common in real-world insurance datasets.

3.1.1. BINARY RELEVANCE

Binary relevance (BR) (see Tsoumakas and Katakis (2008)) decomposes the multilabel classification problem into q independent binary classification tasks, one per label $L_j \in \mathcal{L} = \{L_1, L_2, \dots, L_q\}$. Given an input feature vector $\mathbf{x}_i \in \mathbb{R}^d$, BR assigns a binary output $\hat{y}_{ij} \in \{0, 1\}$ to each label using a dedicated classifier $f_j : \mathbb{R}^d \rightarrow \mathbb{R}$. The prediction is thresholded by a label-specific parameter $\tau_j^* \in [0, 1]$ as follows:

$$\hat{y}_{ij} = \mathbb{I}(f_j(\mathbf{x}_i) \geq \tau_j^*), \quad j \in \{1, \dots, q\}.$$

The complete prediction for observation i is given by $\hat{\mathbf{y}}_i = (\hat{y}_{i1}, \dots, \hat{y}_{iq}) \in \{0, 1\}^q$. While BR is simple and scalable, it assumes independence among labels, potentially limiting performance when label correlations exist.

[Algorithm 1](#) outlines our modified BR approach for addressing the multilabel classification problem. In this framework, each label is treated as an independent binary classification task, allowing the use of any standard classification algorithm as the base learner. In our implementation, we experiment with multiple base classifiers for the function $\mathcal{B}$, including LightGBM (see Ke et al. (2017)) and random forest (see Breiman (2001)). Both classifiers are tree-based models, providing the level of explainability essential for insurance ratemaking. LightGBM uses advanced, scalable algorithms that offer high flexibility, making it particularly well suited for industrial-scale applications. In contrast, while random forest typically demands more computational resources, it consistently delivers strong performance among traditional base learners, especially when applied to large datasets. The flexibility in selecting base classifiers not only allows us to evaluate the accuracy and efficiency of different learning algorithms but also facilitates testing of the dependency structure, while reducing variation attributable to the choice of base model. In addition, validation metric functions F are formally defined and discussed in Appendix B. These measures provide a comprehensive view of model performance by capturing various aspects of classification quality across all labels.

3.1.2. CLASSIFIER CHAINS

Classifier chains (CCs) (see Read et al. (2011)) extend BR by explicitly modeling label dependencies. Let $\sigma = (\sigma_1, \dots, \sigma_q)$ denote a permutation of $\{1, \dots, q\}$ that specifies a fixed label ordering, where σ_j represents the in-

Input: Feature matrix $\mathbf{X}_{\text{train}} \in \mathbb{R}^{r \times d}$; Label matrix $\mathbf{Y}_{\text{train}} \in \{0, 1\}^{r \times q}$; Test feature matrix $\mathbf{X}_{\text{test}} \in \mathbb{R}^{s \times d}$; Test label matrix $\mathbf{Y}_{\text{test}} \in \{0, 1\}^{s \times q}$; Label set $\mathcal{L} = \{L_1, \dots, L_q\}$; Base learner family $\mathcal{B}$ (e.g., LightGBM, Random Forest); Hyperparameter space Θ for the base learner; Threshold space $\mathcal{T}$; The number of cross-validation K ; Loss function; Evaluation metric F

Output: Label-specific thresholds $\tau^* = (\tau_1^*, \dots, \tau_q^*)$; Predicted label matrix $\hat{\mathbf{Y}}_{\text{test}} \in \{0, 1\}^{s \times q}$; Optimal hyperparameter $\theta^* \in \Theta$

```

1 foreach  $(\theta, \tau) \in \Theta \times \mathcal{T}$  do
2   Perform  $K$ -fold cross-validation on  $(\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}})$ 
3   foreach validation fold do
4     Construct dataset:  $\mathcal{D}_j \leftarrow \{(\mathbf{x}_i, y_{ij})\}_{i \in \text{training folds}}$  for each label  $j$ 
5     Train base learner  $f_{j, \theta} \in \mathcal{B}$  on  $\mathcal{D}_j$ 
6     Predict validation scores:  $\mathbf{p}_j^{\text{val}} \leftarrow f_{j, \theta}(\mathbf{X}_{\text{val}})$ 
7     Binarize predictions:  $\hat{\mathbf{y}}_j^{\text{val}} \leftarrow \mathbb{I}(\mathbf{p}_j^{\text{val}} \geq \tau_j)$ 
8     Retrieve validation labels:  $\mathbf{Y}_{\text{val}, j}$ 
9   end
10  Assemble:  $\hat{\mathbf{Y}}_{\text{val}} \leftarrow (\hat{\mathbf{y}}_1^{\text{val}}, \dots, \hat{\mathbf{y}}_q^{\text{val}})$ 
11  Compute cross-validation loss:  $\text{CVLoss}_{\text{fold}} \leftarrow F(\hat{\mathbf{Y}}_{\text{val}}, \mathbf{Y}_{\text{val}})$ 
12 end
13 Select best configuration:  $(\theta^*, \tau^*) \leftarrow \arg \min_{(\theta, \tau)} \frac{1}{K} \sum \text{CVLoss}_{\text{fold}}(\theta, \tau)$ 
14 foreach  $L_j \in \mathcal{L}$  do
15   Train  $f_{j, \theta^*} \leftarrow \mathcal{B}_{\theta^*}$  on  $\{(\mathbf{x}_i, y_{ij})\}_{i=1}^r$ , where  $\mathbf{x}_i \in \mathbf{X}_{\text{train}}, y_{ij} \in \mathbf{Y}_{\text{train}}$ 
16   Set  $\tau_j^* \leftarrow \tau_j$ 
17 end
18 for  $i = 1$  to  $s$  do
19   Retrieve:  $\mathbf{x}_i^* \in \mathbf{X}_{\text{test}}$ 
20   for  $j = 1$  to  $q$  do
21      $\hat{y}_{ij}^* \leftarrow \mathbb{I}(f_{j, \theta^*}(\mathbf{x}_i^*) \geq \tau_j^*)$ 
22   end
23   Set predicted label vector:  $\hat{\mathbf{y}}_i^* \leftarrow (\hat{y}_{i1}^*, \dots, \hat{y}_{iq}^*)$ 
24   Store row  $i$ :  $\hat{\mathbf{Y}}_{\text{test}}[i] \leftarrow \hat{\mathbf{y}}_i^*$ 
25 end
26 return  $(\tau^*, \hat{\mathbf{Y}}_{\text{test}}, \theta^*)$ 

```

Algorithm 1. Binary relevance with joint search for multilabel classification.

dex of the label at position j in the sequence. Since each order defines a distinct sequence of classifiers in the chain, this yields a total of $q!$ permutations, collectively denoted by Σ .

For the first label to predict, σ_1 , the classifier f_{σ_1} takes the feature vector $\mathbf{x}_i$ as input. For each subsequent classifier f_{σ_j} , the input consists of the original feature vector $\mathbf{x}_i$ and the predictions for the labels σ_1 through σ_{j-1} . The prediction for the dimension j in the chain is defined as

$$\hat{y}_{i, \sigma_j} = \mathbb{I}(f_{\sigma_j}(\mathbf{x}_i) \geq \tau_{\sigma_j}^*), \quad j = 1;$$

$$\hat{y}_{i, \sigma_j} = \mathbb{I}(f_{\sigma_j}(\mathbf{x}_i, \hat{y}_{i, \sigma_1}, \dots, \hat{y}_{i, \sigma_{j-1}}) \geq \tau_{\sigma_j}^*),$$

$$j \in \{2, \dots, q\}.$$

[Algorithm 2](#) outlines our modified CC approach for multilabel classification. In this framework, labels are predicted sequentially using a chain of classifiers, where each classifier incorporates previous labels' predictions as additional features, effectively modeling label dependencies. This chaining mechanism enables the model to capture label interdependence. However, its performance is often sensitive to the choice of the label order σ .

3.1.3. MULTILABEL CLASSIFICATION TREES

Multilabel classification trees (MCTs) jointly predict multiple binary labels using a collection of decision trees. Each input $\mathbf{x}_i \in \mathbb{R}^d$ is associated with a vector of binary label $\mathbf{y}_i \in \{0, 1\}^q$, where q is the number of labels.

Unlike BR, which trains one independent model per label, MCTs use a single integrated model to predict all labels simultaneously. Internally, each tree in the ensemble is trained to optimize the average impurity reduction across labels, allowing the model to capture interlabel dependencies. Quan and Valdez (2018) provide a comprehensive discussion of multilabel (multivariate) tree-based models. In this study, we implement the MCT approach using the Python library *RandomForestClassifier*, which ensembles multiple multilabel trees trained on random subsets of the data. [Algorithm 3](#) outlines our modified MCT approach for multilabel classification.

Input: Feature matrix $\mathbf{X}_{\text{train}} \in \mathbb{R}^{r \times d}$; Label matrix $\mathbf{Y}_{\text{train}} \in \{0, 1\}^{r \times q}$; Test feature matrix $\mathbf{X}_{\text{test}} \in \mathbb{R}^{s \times d}$; Test label matrix $\mathbf{Y}_{\text{test}} \in \{0, 1\}^{s \times q}$; Label set $\mathcal{L} = \{L_1, \dots, L_q\}$; Label order space Σ ; Base learner family $\mathcal{B}$ (e.g., LightGBM, random forest); Hyperparameter space Θ for the base learner; Threshold space $\mathcal{T}$; The number of cross-validation K ; Loss function; Evaluation metric F

Output: Optimal label order σ^* ; Label-specific thresholds $\tau^* = (\tau_1^*, \dots, \tau_q^*)$; Predicted label matrix $\hat{\mathbf{Y}}_{\text{test}} \in \{0, 1\}^{s \times q}$; Optimal hyperparameter $\theta^* \in \Theta$

```

1 foreach  $(\sigma, \theta, \tau) \in \Sigma \times \Theta \times \mathcal{T}$  do
2   Perform  $K$ -fold cross-validation on  $(\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}})$ 
3   foreach validation fold do
4     for  $j = 1$  to  $q$  do
5       Construct dataset:  $\mathcal{D}_{\sigma_j}^{\text{fold}} \leftarrow \left\{ (\tilde{\mathbf{x}}_i^{(j)}, y_{i, \sigma_j}) \right\}_{i \in \text{training folds}}$ , where  $\tilde{\mathbf{x}}_i^{(j)} = (\mathbf{x}_i, \hat{y}_{i, \sigma_1}, \dots, \hat{y}_{i, \sigma_{j-1}})$  are
          the augmented features; for  $j = 1$ , we set  $\tilde{\mathbf{x}}_i^{(1)} = \mathbf{x}_i$ .
6       Train base learner  $f_{\sigma_j, \theta} \in \mathcal{B}$  on  $\mathcal{D}_{\sigma_j}^{\text{fold}}$ 
7       Predict validation scores:  $\mathbf{p}_{\sigma_j}^{\text{val}} \leftarrow f_{\sigma_j, \theta}(\tilde{\mathbf{X}}_{\text{val}}^{(j)})$ 
8       Binarize predictions:  $\hat{\mathbf{y}}_{\sigma_j}^{\text{val}} = \mathbb{I}(\mathbf{p}_{\sigma_j}^{\text{val}} \geq \tau_{\sigma_j})$ 
9       Retrieve validation labels:  $\mathbf{Y}_{\text{val}, \sigma_j}$ 
10    end
11    Assemble:  $\hat{\mathbf{Y}}_{\sigma}^{\text{val}} \leftarrow (\hat{\mathbf{y}}_{\sigma_1}^{\text{val}}, \dots, \hat{\mathbf{y}}_{\sigma_q}^{\text{val}})$ 
12    Compute fold loss:  $\text{CVLoss}_{\sigma, \theta, \tau}^{\text{fold}} \leftarrow F(\hat{\mathbf{Y}}_{\sigma}^{\text{val}}, \mathbf{Y}_{\text{val}})$ 
13  end
14  Compute average loss:  $\text{CVLoss}_{\sigma, \theta, \tau} \leftarrow \frac{1}{K} \sum \text{CVLoss}_{\sigma, \theta, \tau}^{\text{fold}}$ 
15 end
16 Select best configuration:  $(\sigma^*, \theta^*, \tau^*) \leftarrow \arg \min_{(\sigma, \theta, \tau)} \text{CVLoss}_{\sigma, \theta, \tau}$ 
17 for  $j = 1$  to  $q$  do
18   Train:  $f_{\sigma_j^*, \theta^*} \leftarrow \mathcal{B}_{\theta^*}$  on  $\mathcal{D}_{\sigma_j^*}^{\text{train}}$ 
19   Set  $\tau_{\sigma_j^*} \leftarrow \tau_{\sigma_j}$ 
20 end
21 for  $i = 1$  to  $s$  do
22   Initialize:  $\tilde{\mathbf{x}}_i^* \in \mathbf{X}_{\text{test}}$ 
23   for  $j = 1$  to  $q$  do
24     Predict:  $p_{\sigma_j^*}^* \leftarrow f_{\sigma_j^*, \theta^*}(\tilde{\mathbf{x}}_i^*)$ 
25     Apply threshold:  $\hat{y}_{i, \sigma_j^*}^* \leftarrow \mathbb{I}(p_{\sigma_j^*}^* \geq \tau_{\sigma_j^*}^*)$ 
26     Augment:  $\tilde{\mathbf{x}}_i^* \leftarrow (\tilde{\mathbf{x}}_i^*, \hat{y}_{i, \sigma_j^*}^*)$ 
27   end
28   Set predicted label vector:  $\hat{\mathbf{y}}_i^* \leftarrow (\hat{y}_{i, \sigma_1^*}^*, \dots, \hat{y}_{i, \sigma_q^*}^*)$ 
29   Store:  $\hat{\mathbf{Y}}_{\text{test}}[i] \leftarrow \hat{\mathbf{y}}_i^*$ 
30 end
31 return  $(\sigma^*, \tau^*, \hat{\mathbf{Y}}_{\text{test}}, \theta^*)$ 

```

Algorithm 2. Classifier chain with joint search for multilabel classification.

3.1.4. CLASSIFICATION PERFORMANCE EVALUATION METRICS

In our study, multilabel classification introduces structural complexities, such as label co-occurrence patterns and imbalanced label distributions, that limit the effectiveness of traditional single-label evaluation metrics. As a result, specialized multilabel performance metrics have been developed to evaluate model performance appropriately.

Among the commonly used evaluation metrics for multilabel classification are the micro-F1, macro-F1, weighted-F1, and sample-F1 scores, each emphasizing different aspects of model behavior. Hinojosa Lee et al. (2024) compare the F1 variants and demonstrate that metric selection can

have a significant effect on evaluation outcomes, particularly under label imbalance. Specifically, the weighted-F1 score further compensates for label imbalance by weighting each label's contribution based on its prevalence in the dataset.

In addition to F1-based metrics, we also use the Jaccard index, which measures set-level similarity, and the Hamming loss, which quantifies the proportion of misclassified labels (see Ganda and Buch (2018)). Collectively, these metrics offer a comprehensive and robust framework for evaluating multilabel classification models under varying label distributions. Appendix B details the evaluation metrics we use in this study.

Input: Feature matrix $\mathbf{X}_{\text{train}} \in \mathbb{R}^{r \times d}$; Label matrix $\mathbf{Y}_{\text{train}} \in \{0, 1\}^{r \times q}$; Test feature matrix $\mathbf{X}_{\text{test}} \in \mathbb{R}^{s \times d}$; Test label matrix $\mathbf{Y}_{\text{test}} \in \{0, 1\}^{s \times q}$; Label set $\mathcal{L} = \{L_1, \dots, L_q\}$; Base learner $\mathcal{B}$ (RandomForestClassifier); Hyperparameter space Θ for the base learner; Threshold space $\mathcal{T}$; The number of cross-validation K ; Loss function; Evaluation metric F

Output: Optimal thresholds $\tau^* = (\tau_1^*, \dots, \tau_q^*)$; Predicted label matrix $\hat{\mathbf{Y}}_{\text{test}} \in \{0, 1\}^{s \times q}$; Optimal hyperparameter $\theta^* \in \Theta$

```

1 foreach  $(\theta, \tau) \in \Theta \times \mathcal{T}$  do
2   Perform  $K$ -fold cross-validation on  $(\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}})$ 
3   foreach validation fold do
4     Train base learner  $f_\theta \in \mathcal{B}$  on training folds
5     Predict validation scores:  $\mathbf{P}^{\text{val}} \leftarrow f_\theta(\mathbf{X}_{\text{val}})$ 
6     Binarize predictions:  $\hat{\mathbf{Y}}^{\text{val}} \leftarrow \mathbb{I}(\mathbf{P}^{\text{val}} \geq \tau)$ 
7     Retrieve validation labels:  $\mathbf{Y}_{\text{val}}$ 
8     Compute cross-validation loss:  $\text{CVLoss}_{\text{fold}} \leftarrow F(\hat{\mathbf{Y}}^{\text{val}}, \mathbf{Y}_{\text{val}})$ 
9   end
10 end
11 Select best configuration:  $(\theta^*, \tau^*) \leftarrow \arg \min_{(\theta, \tau)} \frac{1}{K} \sum \text{CVLoss}_{\text{fold}}(\theta, \tau)$ 
12 Train final model:  $f_{\theta^*} \leftarrow \mathcal{B}_{\theta^*}$  on  $(\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}})$ 
13 Set  $\tau^* \leftarrow (\tau_1^*, \dots, \tau_q^*)$ 
14 for  $i = 1$  to  $s$  do
15   Retrieve:  $\mathbf{x}_i^* \in \mathbf{X}_{\text{test}}$ 
16   Predict all labels jointly:  $\mathbf{p}_i^* \leftarrow f_{\theta^*}(\mathbf{x}_i^*)$ 
17   Apply vectorized thresholding:  $\hat{\mathbf{y}}_i^* \leftarrow \mathbb{I}(\mathbf{p}_i^* \geq \tau^*)$ 
18   Store row  $i$ :  $\hat{\mathbf{Y}}_{\text{test}}[i] \leftarrow \hat{\mathbf{y}}_i^*$ 
19 end
20 return  $(\tau^*, \hat{\mathbf{Y}}_{\text{test}}, \theta^*)$ 

```

Algorithm 3. Multilabel classification trees with joint search.

3.2. MULTI-OUTPUT REGRESSION

Multi-output regression is a supervised learning approach in which each observation is associated with multiple continuous outcomes. By jointly modeling these correlated outcomes, multi-output regression enables more coherent and efficient estimation, enhances predictive accuracy, and supports integrated decision-making, particularly valuable for insurers managing multifaceted risk exposures. Given the structural similarity between multi-output regression and multilabel classification, we omit a separate algorithmic description and instead focus on highlighting the key differences between the two frameworks in the discussion that follows.

3.2.1. MULTI-OUTPUT REGRESSOR

The multi-output regressor (MOR) treats a multi-output regression problem as a set of q independent single-output regression tasks, one for each output dimension $O_j \in \mathcal{O} = \{O_1, O_2, \dots, O_q\}$. Given an input feature vector $\mathbf{x}_i$, the MOR approach constructs a separate regressor $f_j: \mathbb{R}^d \rightarrow \mathbb{R}$ for each output dimension. The regressors are trained independently but in parallel, and the prediction for the dimension j is defined as

$$\hat{z}_{ij} = f_j(\mathbf{x}_i), \quad j \in \{1, \dots, q\}.$$

The complete prediction for observation i is then expressed as $\hat{\mathbf{z}}_i = (\hat{z}_{i1}, \dots, \hat{z}_{iq}) \in \mathbb{R}^q$. This approach, commonly referred to as a problem transformation method, offers flexi-

bility by enabling the use of any standard regression algorithm as the base learner. However, it assumes conditional independence among outputs given the input, which may hinder performance when output dependencies are informative. In our implementation, we test multiple base regressors $\mathcal{B}$, including LightGBM and random forest. We formally define the evaluation metrics used to assess model performance across all dimensions in Appendix C.

3.2.2. REGRESSOR CHAIN

The regressor chain (RC) (see Spyromitros-Xioufis et al. (2012)) is inspired by chain-based strategies originally developed for multilabel classification (see Read et al. (2011)), which incorporate sequential dependency modeling among output variables. Similarly to CCs in the multilabel setting before, RC defines a permutation $\sigma = (\sigma_1, \dots, \sigma_q)$ of the output indices $\{1, \dots, q\}$, determining the fixed ordering of output dimensions within the chain. Except for the regressor f_{σ_1} , which corresponds to the first label σ_1 and requires only the feature vector $\mathbf{x}_i$ as input, for each succeeding regressor f_{σ_j} , the input is formed by augmenting the original feature vector $\mathbf{x}_i$ with the predicted values of the preceding outputs in the chain, and it returns a prediction for the output dimension indexed by σ_j :

$$\begin{aligned} \hat{z}_{i, \sigma_j} &= f_{\sigma_j}(\mathbf{x}_i), \quad j = 1; \\ \hat{z}_{i, \sigma_j} &= f_{\sigma_j}(\mathbf{x}_i, \hat{z}_{i, \sigma_1}, \dots, \hat{z}_{i, \sigma_{j-1}}), \\ &\quad j \in \{2, \dots, q\}. \end{aligned}$$

This chaining mechanism also allows the model to take advantage of the dependencies between the outputs. However, similar to its classification counterpart, RC is sensitive to the specified output ordering σ . Furthermore, in scenarios where output dependencies are weak or negligible, the chaining mechanism may even introduce instability in model performance or additional computational burdens.

3.2.3. MULTI-OUTPUT REGRESSION TREES

Multi-output regression trees (MRTs) extend traditional decision trees to simultaneously predict multiple continuous outputs. Rather than training separate models for each output, MRTs construct a unified tree structure that partitions the input space to minimize a multivariate loss function. This formulation allows the model to account for correlations among outputs, and for each input $\mathbf{x}_i \in \mathbb{R}^d$, the model estimates a vector $\mathbf{z}_i \in \mathbb{R}^q$.

In this study, we implement the MRT approach with *RandomForestRegressor* to construct a unified ensemble model. However, as the number of output dimensions increases ($q \gg 1$), evaluating multivariate splits becomes increasingly computationally intensive, which may hinder scalability.

3.2.4. REGRESSION PERFORMANCE EVALUATION METRICS

Compared with single-output regression tasks, where standard metrics such as mean squared error (MSE) or R-squared (R^2) are sufficient, multi-output regression requires evaluation metrics that account for the joint behavior of multiple continuous outputs. In this study, we evaluate model performance using the following six metrics: average MSE (aMSE), average RMSE (aRMSE), average relative RMSE (aRRMSE), average correlation coefficient (aCC), and global Euclidean distance (EU_DIST), as described in Appendix C. These metrics collectively capture different dimensions of model performance, including error magnitude, scale-adjusted accuracy, and joint prediction consistency (see Borchani et al. (2015)).

4. DATA-DRIVEN ANALYSIS AND FINDINGS

In our empirical study, the classification task corresponds to capturing the occurrence of different types of cyber incidents, analogous to claim occurrence modeling in the insurance industry. Meanwhile, the regression task focuses on estimating the frequency or expected count of incidents, consistent with frequency modeling commonly used to assess actuarial pricing.

4.1. CONVENTIONAL DATA VERSUS INSURTECH-ENRICHED DATA

To evaluate the efficacy of entity-specific organizational features derived from insurtech on both cyber incident occurrence and frequency, we perform a comparative analysis using two datasets with varying features but the same label: an existing conventional cyber incident dataset (D1) con-

taining only the most commonly used rating factors, including industry classification and annual revenue, and an insurtech-enriched dataset (D2) that incorporates additional external entity-specific organizational features obtained from various platforms (with details provided in Section 2.2). These enriched features provide an entity-specific representation of an organization’s cyber risk posture. Note that the features in D1 are a subset of those in D2.

The conventional dataset (D1) is evaluated using three representative models for each task, while the enriched dataset (D2) is evaluated using five model configurations to enable a broader assessment of both occurrence and frequency within the enriched data setting. Specifically, for classification, we implement BR using random forest and LightGBM classifiers extended via *MultiOutputClassifier* (RF_MOC and LGBM_MOC) and implement CC using the same base classifiers (namely, RF_CC and LGBM_CC). We also incorporate MCTs, denoted as MLRF, which leverage tree ensembles to simultaneously predict multiple labels within a unified classification framework. For regression, we implement MOR using random forest and LightGBM regressors wrapped with *MultiOutputRegressor* (namely, RF_MOR and LGBM_MOR), where each output is modeled independently without capturing interdependencies. To explore potential dependencies among outputs, we also implement RC with the same base regressors (namely, RF_RC and LGBM_RC). Additionally, we use MRTs, termed MORE, to jointly predict all output variables.

To facilitate a comprehensive comparison across datasets and model architectures, we construct comparative heatmaps to visualize the model performances across both classification and regression tasks. In our experiments, we evaluate model performance using two separate sets of evaluation metrics mentioned in Appendix B and Appendix C. To ensure consistency and facilitate comparison, evaluation metric values within each task are normalized across models to generate the heatmap’s color gradient (darker blue tones reflect better model performance, in contrast to red tones indicating weaker results), while the original evaluation metric values are displayed within the heatmap cells to retain precise quantitative detail. Therefore, each row of the heatmap matrix represents a specific model, while each column corresponds to an evaluation metric, allowing for a systematic comparison of model performance across various architectures and dataset configurations.

For cyber incident occurrence evaluation, we use six metrics: weighted-F1, macro-F1, micro-F1, and sample-F1 scores, as well as the Jaccard index and Hamming loss. In our analysis, we focus on weighted-F1 and Hamming loss as the most informative indicators of model performance (with details provided in Section 3.1.4). As shown in [Figure 1](#), models trained on D2 consistently outperform those trained on D1 across both weighted-F1 and Jaccard index. Weighted-F1 scores on D2 reach values above 0.88 in the training set and over 0.64 in the test set, while the same metrics on D1 remain below 0.63. Additionally, the Jaccard index improves from an average of less than 0.61 on D1 to consistently exceeding 0.61 on D2. These re-

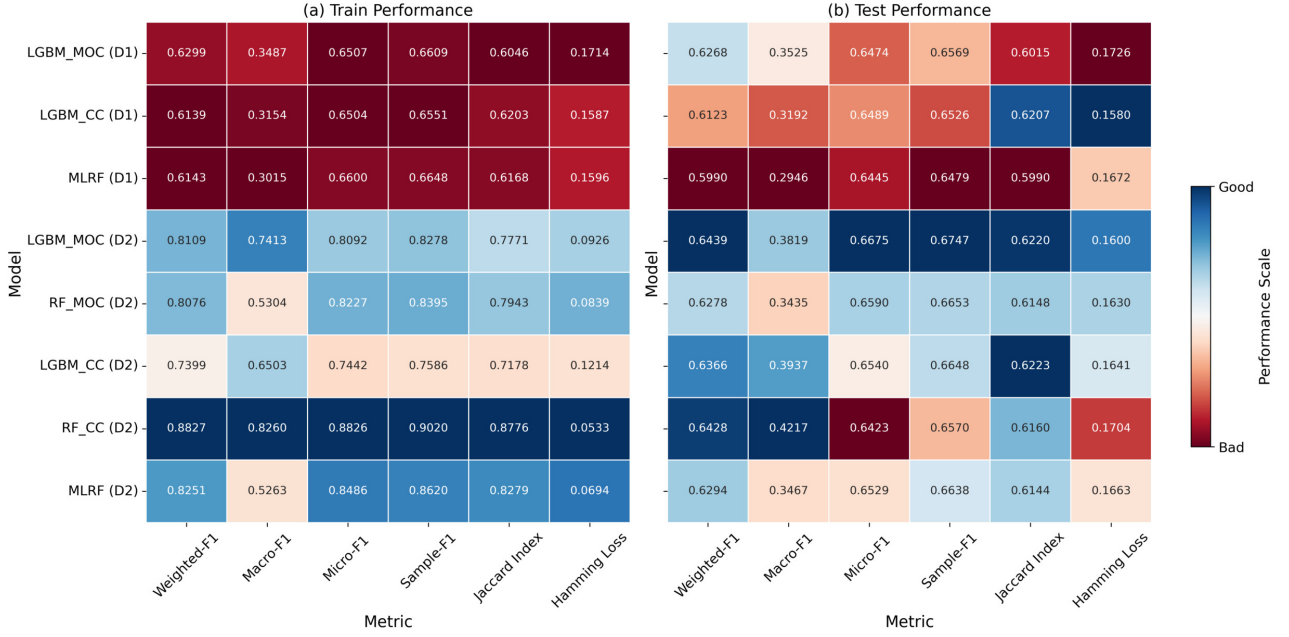


Figure 1. Multilabel classification model performance.

sults demonstrate that entity-specific organizational features from insurtech capture critical information that improves the model’s ability to correctly identify multiple types of incidents, leading to more accurate and reliable performance.

For cyber incident frequency evaluation, we assess model performance using aMSE, aRMSE, aCC, aRRMSE, and EU_DIST, as summarized in Section 3.2.4. Overall, the enriched dataset (D2) provides substantial improvements in frequency estimation performance, particularly for models without dependency structures, as illustrated in Figure 2. Specifically, models trained on D2 consistently achieve lower aRMSE and aMSE values, with aRMSE decreasing from approximately 0.36 on D1 to as low as 0.28 on D2 in training, and the corresponding test set values follow a similar trend. Likewise, the aMSE values on D2 decrease below 0.11 in the train set and around 0.16 in the test set, compared with the higher values observed on D1. However, these improvements are not consistent across all model types, as certain dependency-based models, such as RC, exhibit degraded performance under D2. This highlights that misspecification in the dependency structure, especially when using high-dimensional entity-specific organizational features, can lead to overfitting on the training data. As a result, even with insurtech-enriched data, frequency estimates may be adversely affected.

These findings collectively demonstrate that insurtech-enriched entity-specific organizational features play a pivotal role in improving occurrence modeling and frequency estimation by providing more informative and context-specific inputs.

4.2. METRIC-SPECIFIC MODEL PERFORMANCE

On the insurtech-enriched dataset (D2), LGBM_MOC achieves the highest weighted-F1 score (0.6439), demon-

strating superior overall classification performance after adjusting for label imbalance. This indicates that the model performs effectively across the entire label distribution, including both frequent and infrequent classes. RF_CC attains a very similar weighted-F1 score (0.6428); however, it exhibits the highest Hamming loss (0.1704) among the D2 models, reflecting a greater average number of label-wise prediction errors per instance. In contrast, LGBM_MOC attains the second-lowest Hamming loss (0.1600), suggesting more conservative and precise label-wise predictions. This trade-off highlights the importance of evaluating multilabel models across complementary metrics: Weighted-F1 reflects performance relative to class distribution, whereas Hamming loss penalizes scattered misclassifications and overpredictions.

For frequency estimation, LGBM_MOR achieves the best performance across all metrics. On D2, this model records an aRMSE of 0.2830 (train) and 0.3477 (test), along with an aMSE of 0.1008 (train) and 0.1533 (test), representing substantial reductions in average prediction errors compared to all D1 counterparts. The model also attains aCC values of 0.7128 (train) and 0.3623 (test), indicating improved alignment between predicted and observed incident frequencies. Similarly, the MORF model on D2 demonstrates lower aRMSE and aMSE values compared with its D1 version, confirming that entity-specific organizational insurtech-empowered features significantly enhance accuracy for models without explicit dependency structures. However, RC models show the opposite trend: Both RF_RC and LGBM_RC experience an increase in aRMSE (up to 0.3732 in the test set) and a decrease in aCC (as low as 0.2336) on D2, indicating both higher prediction errors and weaker alignment with observed frequencies. Although LGBM_MOR shows consistently superior results across all metrics, the performance ranking of other models differs across different metrics.

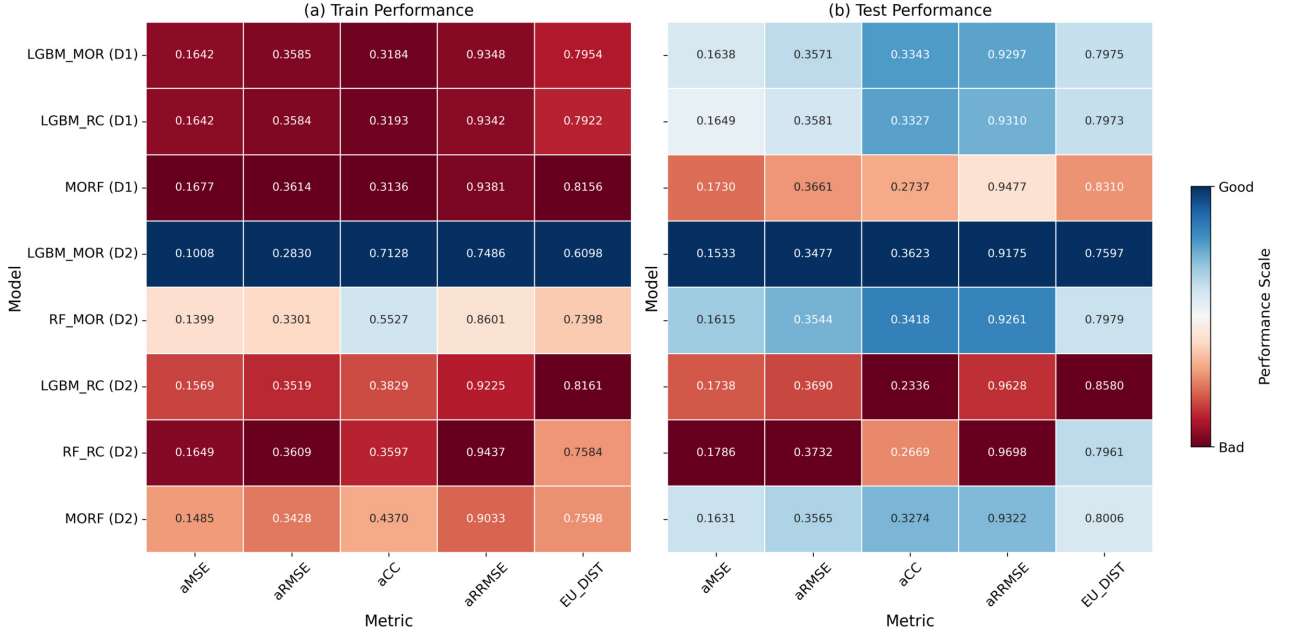


Figure 2. Multi-output regression model performance.

This lack of consistency illustrates that no single metric fully reflects model quality, highlighting the importance of multimetric evaluation, especially under insurtech-enriched feature settings.

4.3. MODEL ARCHITECTURE

Although classification model architectures differ in their capacity to capture interlabel dependencies, their overall impact on model performance is relatively limited within the context of our study. Architectures such as CC (e.g., RF_CC, LGBM_CC) and MLRF are explicitly constructed to capture interlabel dependencies and facilitate joint optimization over multiple output variables. Nonetheless, empirical results reveal that these structurally sophisticated models do not consistently yield superior model performance compared with simpler alternatives. Notably, LGBM_MOC, a relatively straightforward model based on binary relevance with multi-output adaptation, attains the highest weighted-F1 score (0.6439) on D2, outperforming LGBM_CC (0.6366) and MLRF (0.6294).

In addition, for frequency estimation tasks, the dependency architecture has an inferior effect. Our results reveal that RC models exhibit instability under D2, where enriched entity-specific organizational insurtech features increase data complexity. RC models explicitly capture sequential output dependencies, incorporating the prediction order as a tunable hyperparameter. While, in theory, optimizing this order can improve model performance, the factorial growth of possible permutations with increasing label dimensionality makes exhaustive tuning computationally expensive. Empirically, both RF_RC and LGBM_RC demonstrate elevated aMSE and aRMSE, reaching up to 0.1786 and 0.3732 on the test set, alongside decreased CC values as low as 0.2336. These outcomes suggest that the interaction between insurtech-enriched features and dependency

structures introduces additional system complexity, magnifying the sensitivity to output ordering and increasing the susceptibility to overfitting. Alternatively, dependency-free models such as LGBM_MOR and MORF consistently achieve lower errors and greater prediction stability under D2, demonstrating the advantages of simpler architectures in effectively leveraging enriched entity-specific organizational features for reliable frequency estimation.

Therefore, in our cyber incident dataset, dependencies between different types of cyber incidents appear to be weak or nonexistent. Nevertheless, model architecture remains a critical consideration, as effectively leveraging enriched data requires selecting structures that balance model complexity with stability, especially in high-dimensional feature spaces.

5. FEATURE IMPORTANCE

With over 500 features in the insurtech-enriched dataset, interpreting the factors driving model performance becomes challenging. To enhance the interpretability of the results, we identify the top 20 most influential features that contribute to the model performance of the multilabel classification and multi-output regression models described in Section 3.1 and Section 3.2. To achieve this, we apply several widely used feature importance techniques, aiming to uncover key features and provide insights into potential cyber risk factors captured by the insurtech data.

5.1. FEATURE IMPORTANCE TECHNIQUES

It is important to emphasize that feature importance techniques provide insight into how features influence a model's predictions, but they do not imply causal relationships. The observed associations reflect the model's in-

ternal mechanics rather than underlying cause-and-effect dynamics. Establishing causality would require a separate, rigorous analysis using formal causal inference methods, supported by domain expertise and possibly experimental or quasi-experimental designs.

Nonetheless, feature importance techniques serve as valuable tools for interpreting complex ML models, offering a window into the so-called “black box.” However, each technique has its own limitations, such as sensitivity to feature correlation, model bias, or instability under resampling, which can affect interpretability. To mitigate those concerns and improve the robustness of our findings, we apply multiple feature importance techniques: impurity-based importance, permutation importance, and SHapley Additive exPlanations (SHAP). This ensemble approach enables a more comprehensive and reliable identification of the key features driving the model’s predictive performance.

5.1.1. IMPURITY-BASED FEATURE IMPORTANCE

Impurity-based feature importance, often referred to as mean decrease impurity, introduced by Breiman (2002), quantifies the influence of each input feature by aggregating its contributions to reductions in a chosen impurity criterion during the training process. For example, in random forest classification, the two predominant impurity criteria are the Gini index (see Lerman and Yitzhaki (1984)) and information gain (see Kent (1983)). The Gini index, widely adopted because of its computational efficiency, measures the probability of misclassification and reflects class heterogeneity. Alternatively, information gain, derived from entropy, serves as a more discriminative impurity measure by quantifying the expected reduction in uncertainty after a split. Compared with the Gini index, information gain often exhibits greater sensitivity to informative features in high-dimensional spaces, where many features may be irrelevant or redundant. By assigning minimal importance to uninformative features, it encourages sparsity in feature selection. However, its higher computational cost can limit scalability in large ensembles. Unlike classification trees that assess class purity, regression trees assess impurity in terms of prediction error. Specifically, reductions in MSE are used to evaluate split quality, and each decrease in output variance is attributed to the corresponding partitioning feature. In LightGBM, feature importance is frequently assessed using gain and split count. Gain measures the total reduction in the loss function resulting from splits involving a given feature, while split count indicates how often the feature is used for partitioning across all decision nodes. Unlike classical information gain, which is based on entropy reduction in classification tasks, LightGBM gain derives from improvements in the model-specific loss function (e.g., MSE or log loss), making it applicable to both regression and classification.

5.1.2. PERMUTATION FEATURE IMPORTANCE

Permutation feature importance, introduced by Breiman (2001), and formalized by Fisher et al. (2019), is a method

to evaluate the contribution of a feature by measuring the change in model performance after randomly permuting that feature’s values. This approach maintains the original model and prediction process but introduces noise to one feature at a time to observe the effect on predictive accuracy. The difference in the evaluation metric before and after permutation indicates the importance of the feature. The method is model-agnostic, applicable to both classification and regression tasks, and supports a consistent assessment framework across different algorithms. It is particularly useful for interpreting complex models such as ensembles or neural networks. However, it may be sensitive to feature correlation, as highly correlated features can mask each other’s contribution during permutation.

5.1.3. SHAPLEY ADDITIVE EXPLANATIONS FEATURE IMPORTANCE

SHAP, introduced by Lundberg and Lee (2017), interprets model predictions by computing Shapley values derived from cooperative game theory. It explains each individual prediction by assigning every feature a local importance value based on its marginal contribution, averaged over all possible subsets of input features. This method ensures a fair and theoretically consistent allocation of feature contributions, satisfying key properties such as local accuracy, consistency, and missingness. To assess global feature importance, SHAP aggregates the absolute Shapley values across all observations, yielding a comprehensive ranking of features by their average contribution to model output. Compared with impurity-based and permutation-based approaches, SHAP captures interaction effects and the direction of influence (positive or negative), though it often requires greater computational resources. For tree-based models such as random forest and LightGBM, SHAP values can be computed efficiently using TreeExplainer (see Lundberg et al. (2020)), a popular explainer tailored to ensemble tree models. TreeExplainer leverages the structure of decision trees to reduce the computational complexity from exponential to polynomial time, enabling exact and fast computation of Shapley values.

5.2. ENTITY-LEVEL FEATURE CONTRIBUTIONS TO CYBER INCIDENT OCCURRENCE AND FREQUENCY

To examine how entity-specific organizational features influence the occurrence of cyber incident types, we analyze feature importance across five model configurations: LGBM_MOC, LGBM_CC, RF_MOC, RF_CC, and MLRF. In addition, we extend this analysis to frequency estimation by evaluating feature importance across five regression model configurations: LGBM_MOR, LGBM_RC, RF_MOR, RF_RC, and MORF. For both tasks, we apply three different feature importance techniques—impurity-based importance, permutation importance, and SHAP—to provide a comprehensive evaluation of influential entity-specific organizational features.

All analyses are conducted using the insurtech-enriched dataset (D2), which incorporates external entity-specific organizational features beyond the conventional cyber in-

cident dataset. Figure 3 reports the log-transformed feature importance scores for the classification task. Each classification model produces multiple importance measures (e.g., impurity-based, permutation-based, SHAP-based). The features are ranked using their average importance across these measures, and the top 20 features under this ranking are visualized. The heatmap displays the importance values of these features across all individual measures. Figure 4 reports how consistently each feature is selected across the different importance measures within each classification model. A higher count reflects greater cross-measure stability, indicating that the feature is reliably identified as important regardless of the specific importance metric used. Analogous summaries for the regression models are presented in Figure 5 and Figure 6. In both heatmaps, darker color intensity indicates higher importance or greater stability.

Building on the four heatmaps described above, we find that several features consistently stand out as influential, regardless of the underlying model architecture or the specific importance metric used. This consistency suggests that certain entity-specific organizational features play a systematically important role in capturing both incident occurrence and frequency. To provide a clearer and more accessible comparison of these recurring features, we extract the top 10 features based on the combined evidence from all four heatmaps. A consolidated summary of those results is presented in Appendix E, Table 5. That table offers a direct, side-by-side overview of the dominant features, integrating information from all models and importance measures to highlight features that exhibit strong and persistent informative value across tasks.

Geographic indicators—a subset of *Business Information* features in Table 4—consistently demonstrate predictive value across both tasks, though specific features differ by modeling objective. For occurrence modeling, address features (i.e., *ADDRESS_state_east*, *ADDRESS_state_west*, and *ADDRESS_street_type_St*) rank highly, suggesting that regional location and address structure influence the likelihood of incident occurrence, potentially reflecting regulatory variations or localized vulnerability. That is, some states implement more stringent security incident notification laws, increasing the likelihood of incidents being reported. Similarly, for frequency estimation, geographic coordinates (i.e., *ADDRESS_coordinates_lon* and *ADDRESS_coordinates_lat*) emerge as key features, capturing spatial heterogeneity in incident frequency patterns, likely related to geographic disparities in cyber exposure or reporting practices.

Moreover, industry composition emerges as an important determinant for both occurrence and frequency estimation tasks. Sector-specific classification features (i.e., features beginning with *CLASSIFICATION_segment*) highlight that the industry sector composition affects not only the likelihood of incident occurrence but also its frequency. This result is consistent with findings in existing literature and current industry practices that regard industry classification as one of the most important rating factors. While broad sector-level classification features already rank

highly across models, more granular measures of industry composition provide additional explanatory power. A key feature is the indicator for whether a firm is engaged in office and administrative support services (i.e., *CLASSIFICATION_segment_details_proportion_segment_office_administrative_support_service*), which belongs to the *Classification: Segment Proportion* category in Table 4. Its domain influence suggests that firms engaged more heavily in such sectors may be exposed to heightened cyber risk frequency, potentially due to their reliance on extensive administrative processes, routine handling of sensitive information, and higher volumes of digital interactions. This highlights that beyond sector membership, the inclusion of specific industry activities serves as a critical determinant of cyber incident risk.

Meanwhile, the number of business risk characteristics (i.e., *CHARACTERISTICS_business_characteristics_bc_count*) is consistently prioritized in the occurrence and frequency modeling. This feature belongs to the broader *Risk Characteristics* category mentioned in Table 4 and its prioritization suggests that organizations with greater operational breadth or business diversity face a higher likelihood of experiencing cyber incidents, likely due to increased complexity or broader exposure surfaces.

While many influential features overlap between occurrence and frequency modeling, distinct task-specific patterns are also observed. For frequency estimation, features from the *Proximity Score* category (i.e., *CARPE_SCORES_proximity_scores_traffic_mode_distance_version_1_0*) emerge among the most influential features. Proximity scores capture how closely an organization is situated relative to surrounding activity or foot-traffic intensity. Their strong importance implies that firms located in more active or densely connected environments may experience higher incident frequency given their greater external exposure. This observation is also consistent with the high-importance geographic indicators we discussed earlier. Reputation-related features also show clear relevance in the frequency models. Features in the *Index* category, such as *REVIEW_SCORE_average_score*, consistently contribute to explaining incident frequency, indicating that external customer-facing evaluations carry information about firms' underlying risk posture. In contrast, features in the *Proximity* category play a more prominent role in the occurrence models. For example, *CARPE_SCORES_negative_keyword_score_keyword_count* is repeatedly selected as an important feature for incident occurrence. Higher levels of negative online sentiment appear to be associated with an elevated probability of cyber incidents, suggesting that firms experiencing negative external signals may also face heightened operational exposure.

In addition to these entity-specific organizational features, the incident year (i.e., *ACCIDENT_YEAR*), which does not originate from the insurtech feature set but instead comes directly from the incident dataset, also emerges as one of the most influential features in both modeling tasks. Its strong predictive value likely reflects broader temporal dynamics, including shifts in the threat landscape, changes

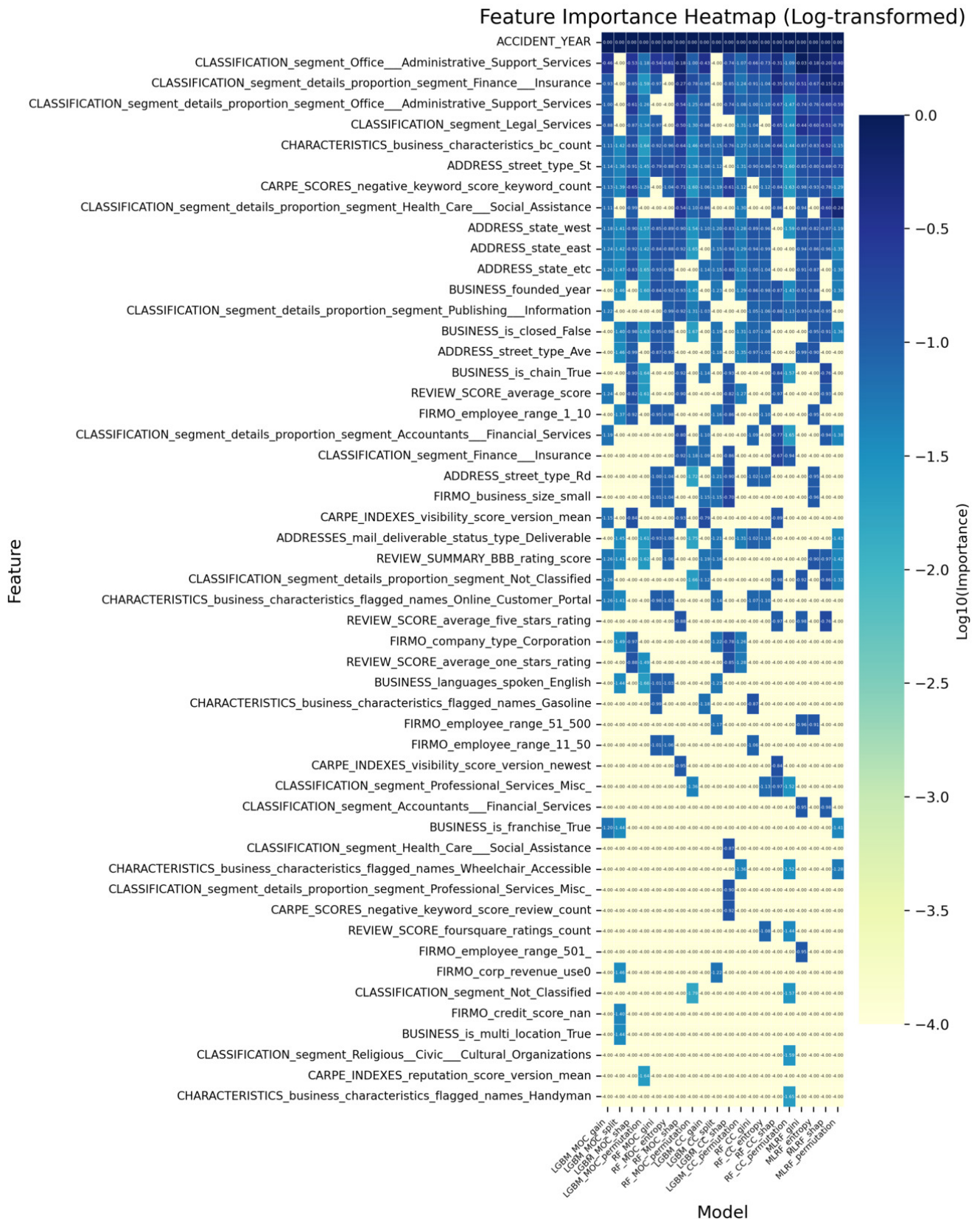


Figure 3. Log-transformed feature importance scores across various classification models.



Figure 4. Count of top important feature appearance across various classification models.

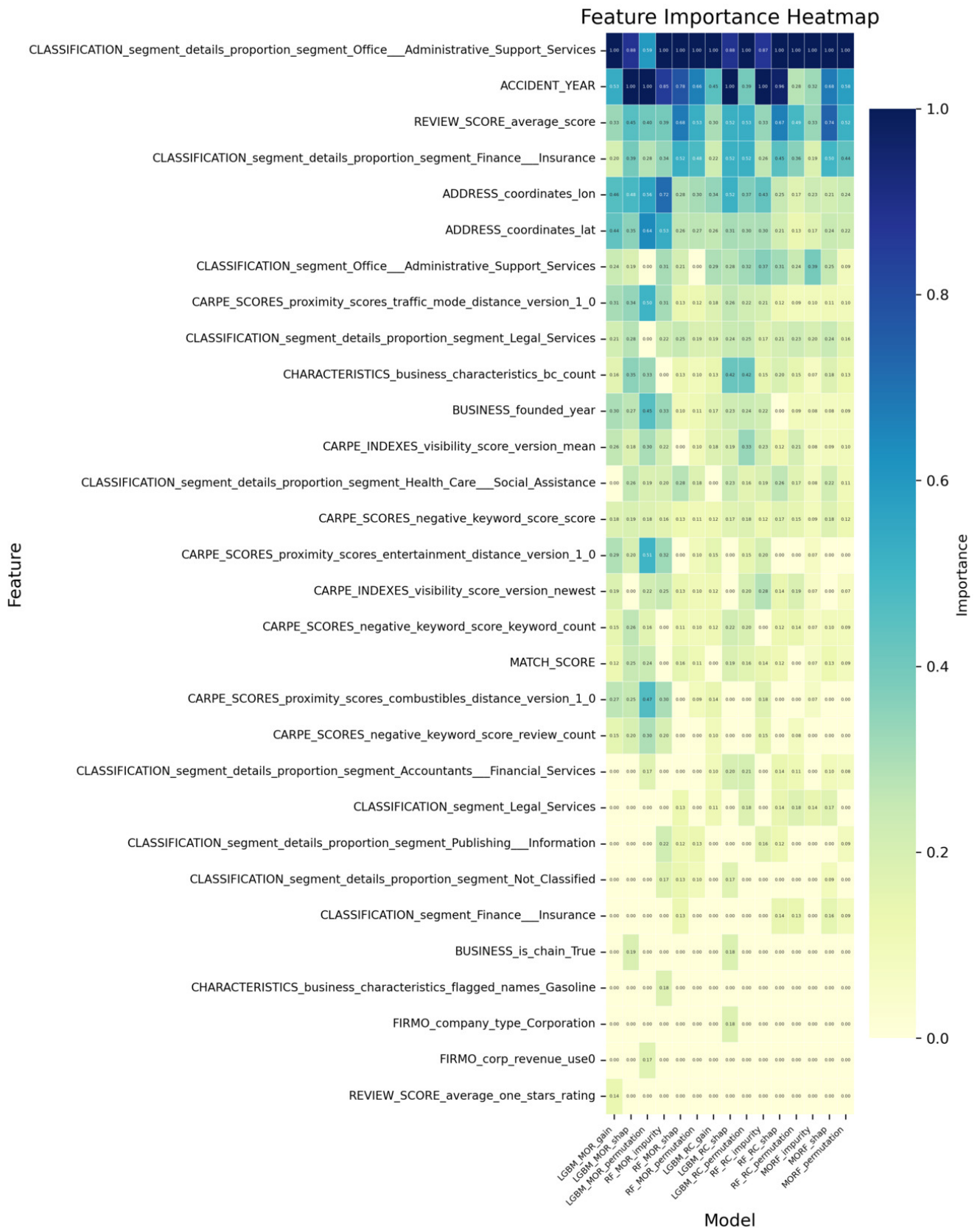


Figure 5. Feature importance scores across various regression models.



Figure 6. Count of top important feature appearance across various regression models.

in reporting practices, and evolving regulatory requirements across years.

In summary, the consistent convergence across multiple importance measures indicates that many of the most influential features are concentrated in a few categories of entity-specific organizational features—*Business Information* (e.g., geographic indicators), *Classification* (e.g., segment proportions), and *Proximity Score*. Within these categories, specific features repeatedly emerge as top-ranked features for both the occurrence and frequency of cyber incidents. This supports our central hypothesis that enriching sparse cyber incident data with entity-specific organizational features enables more granular and interpretable models for cyber risk.

6. DISCUSSION AND CONCLUSION

The modeling results suffice to answer the questions of interest raised in Section 1. Entity-specific organizational features, in addition to industry classification and revenue, can offer substantial performance improvement in capturing the occurrence and frequency of cyber incidents. In particular, reputation-related features are found to play a pivotal role in affecting a firm’s cyber risk, and this aspect has received limited attention and empirical validation in existing research. In addition, with the enriched dataset, the classification models demonstrate strong model performance in identifying the occurrences of cyber incidents. However, despite improvements driven by entity-specific organizational features, accurately predicting the frequencies of various types of incidents remains a challenging task. Lastly, we find little evidence for dependence among either occurrences or frequencies of different types of cyber incidents at the entity level. That is, once firm-specific characteristics are taken into account, the occurrences and frequencies of different incident types are likely to be conditionally independent, and the occurrence or frequency of one incident type does not, in itself, provide predictive power for other types. The observed conditional independence is also supported by the almost negligible partial correlations of occurrences and frequencies across incident types after removing the effect of firm-specific characteristics (see Appendix F). This finding, however, does not address the unconditional dependence or independence among incident types.

To understand the performance gap between the frequency and occurrence models, we turn to the construction of our modeling dataset and the empirical characteristics of the cyber incident data. Two key factors drive the discrepancy. First, occurrence modeling is a binary classification task concerned only with identifying whether at least one incident occurs in a given firm-year. In contrast, frequency modeling requires predicting the number of incidents, which is significantly more complex. Cyber incident counts are highly skewed and dominated by zeros, leaving limited usable variation for estimating the conditional distribution of incident frequency. Second, the empirical distribution of cyber incidents is extremely imbalanced across incident types as well. As illustrated in Appendix G, [Figure](#)

[7a](#), the majority of occurrence records arise from *Data Breach* and *Privacy Violation*, whereas categories such as *Extortion/Fraud*, *IT Error*, and *Other* appear only sporadically. A similar pattern emerges for incident frequency. As shown in Appendix G, [Figure 7b](#), the total counts of incidents (i.e., aggregated frequency) exhibit a heavily right-skewed distribution, again dominated by *Privacy Violation* and *Data Breach*. This combination of modeling data creates imbalance and heavy tails, substantially reducing the effectiveness of frequency modeling, making it statistically more difficult than occurrence prediction.

Some limitations of this study are as follows and may motivate future research. For one, constrained by the availability of cyber loss data in the public domain, the study focuses only on the occurrence and frequency of cyber incidents. Whereas some empirical studies have examined the magnitude of cyber losses, none have conducted severity predictions using an extensive set of firm-specific features. Such severity models can potentially be established using proprietary datasets, such as the cyber loss data from Advisen and the claims data managed by cyber insurance providers. Moreover, beyond the insurtech-enhanced features explored in this study, there remains a significant opportunity for future enhancement in identifying and incorporating cyber risk factors. One promising direction is the integration of penetration test data, which leverages engineering domain expertise to replicate real-world attack environments. Such simulated scenarios could provide a richer understanding of system vulnerabilities and organizational response behaviors, offering valuable predictive signals that are difficult to capture through observational data alone. Incorporating such data could further strengthen the robustness and realism of cyber risk models.

Given the evolving and dynamic nature of cyber threats, it is increasingly urgent to develop a quantitative underwriting framework grounded in robust predictive analytics and sound risk management principles. By integrating high-frequency entity-specific data into underwriting models, insurers can improve risk differentiation, enhance pricing accuracy, and promote a more resilient cyber insurance market.

ACKNOWLEDGMENTS

Zhiyu Quan and Linfeng Zhang are supported by a 2024 Casualty Actuarial Society individual research grant. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Casualty Actuarial Society. Zhiyu Quan would like to thank the Gies College of Business at the University of Illinois Urbana-Champaign for its support of this work through the ORMIR Faculty Scholars Program.

Submitted: July 10, 2025 EDT. Accepted: December 09, 2025 EDT. Published: February 20, 2026 EDT.

REFERENCES

- Biener, Christian, Martin Eling, and Jan Hendrik Wirfs. 2015. "Insurability of Cyber Risk: An Empirical Analysis." *The Geneva Papers on Risk and Insurance—Issues and Practice* 40 (1): 131–58. <https://doi.org/10.1057/gpp.2014.19>.
- Borchani, Hanen, Gherardo Varando, Concha Bielza, and Pedro Larranaga. 2015. "A Survey on Multi-Output Regression." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5 (5): 216–33. <https://doi.org/10.1002/widm.1157>.
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Breiman, Leo. 2002. *Manual on Setting Up, Using, and Understanding Random Forests V3.1*. Statistics Department, University of California Berkeley.
- Cebula, James J., Mary E. Popeck, and Lisa R. Young. 2014. *A Taxonomy of Operational Cyber Security Risks Version 2*. Carnegie-Mellon University. <https://doi.org/10.21236/ADA609863>.
- Chong, Wing Fung, Daniël Linders, Zhiyu Quan, and Linfeng Zhang. 2025. "Incident-Specific Cyber Insurance." *ASTIN Bulletin* 55 (2): 395–425. <https://doi.org/10.1017/asb.2025.9>.
- Doe, Dissent. 2025. "DataBreaches.Net." <https://databreaches.net/>.
- Eling, Martin, and Kwangmin Jung. 2018. "Copula Approaches for Modeling Cross-Sectional Dependence of Data Breach Losses." *Insurance: Mathematics and Economics* 82: 167–80. <https://doi.org/10.1016/j.insmatheco.2018.07.003>.
- Eling, Martin, and Jan Wirfs. 2019. "What Are the Actual Costs of Cyber Risk Events?" *European Journal of Operational Research* 272 (3): 1109–19. <https://doi.org/10.1016/j.ejor.2018.07.021>.
- Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *Journal of Machine Learning Research* 20 (177): 1–81.
- Ganda, Dhatri, and Rachana Buch. 2018. "A Survey on Multi Label Classification." *Recent Trends in Programming Languages* 5 (1): 19–23.
- Harry, Charles, and Nancy Gallagher. 2018. "Classifying Cyber Events." *Journal of Information Warfare* 17 (3): 17–31.
- Hinojosa Lee, Maria Cristina, Johan Braet, and Johan Springael. 2024. "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores." *Applied Sciences* 14 (2): 9863.
- Hunt, Troy, Charlotte Hunt, and Stefán Jökull Sigurðarson. 2025. "Who's Been Pwned." Have I Been Pwned. <https://haveibeenpwned.com/PwnedWebsites>.
- IBM. 2024. *Cost of a Data Breach 2024*. <https://www.ibm.com/reports/data-breach>.
- Ke, Guolin, Qi Meng, Thomas Finley, et al. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30.
- Kent, John T. 1983. "Information Gain and a General Measure of Correlation." *Biometrika* 70 (1): 163–73. <https://doi.org/10.1093/biomet/70.1.163>.
- Kesan, Jay P., and Linfeng Zhang. 2020. "When Is a Cyber Incident Likely to Be Litigated and How Much Will It Cost? An Empirical Study." *Connecticut Insurance Law Journal* 27 (2): 529–80.
- Kesan, Jay P., and Linfeng Zhang. 2021. "An Empirical Investigation of the Relationship Between Local Government Budgets, IT Expenditures, and Cyber Losses." *IEEE Transactions on Emerging Topics in Computing* 9 (2): 582–96. <https://doi.org/10.1109/TETC.2019.2915098>.
- Ko, Ryan, Elinor Tsen, and Sergeja Slapnicar. 2020. "Dataset of Data Breaches and Ransomware Attacks over 15 Years from 2004." University of Queensland. <https://doi.org/10.14264/dfc5027>.
- Lerman, Robert I., and Shlomo Yitzhaki. 1984. "A Note on the Calculation and Interpretation of the Gini Index." *Economics Letters* 15 (3–4): 363–68. [https://doi.org/10.1016/0165-1765\(84\)90126-5](https://doi.org/10.1016/0165-1765(84)90126-5).
- Lundberg, Scott M., Gabriel Erion, Hugh Chen, et al. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30.
- Nurse, Jason R. C., Louise Axon, Arnau Erola, Ioannis Agraftiotis, Michael Goldsmith, and Sadie Creese. 2020. "The Data That Drives Cyber Insurance: A Study into the Underwriting and Claims Processes." In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*. IEEE. <https://doi.org/10.1109/CyberSA49311.2020.9139703>.
- Palsson, Kjartan, Steinn Gudmundsson, and Sachin Shetty. 2020. "Analysis of the Impact of Cyber Events for Cyber Insurance." *The Geneva Papers on Risk and Insurance—Issues and Practice* 45 (4): 564–79. <https://doi.org/10.1057/s41288-020-00171-w>.
- Privacy Rights Clearinghouse. 2025. "Data Breach Chronology." <https://privacyrights.org/data-breaches>.
- Quan, Zhiyu, Changyue Hu, Panyi Dong, and Emiliano A. Valdez. 2024. "Improving Business Insurance Loss Models by Leveraging InsurTech Innovation." *North American Actuarial Journal* 29 (2): 247–74. <https://doi.org/10.1080/10920277.2024.2400648>.
- Quan, Zhiyu, and Emiliano A. Valdez. 2018. "Predictive Analytics of Insurance Claims Using Multi-Variate Decision Trees." *Dependence Modeling* 6 (1): 377–407. <https://doi.org/10.1515/demo-2018-0022>.

- Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. "Classifier Chains for Multi-Label Classification." *Machine Learning* 85: 333–59. <https://doi.org/10.1007/s10994-011-5256-5>.
- Rege, Aunshul, Rachel Bleiman, et al. 2023. "A Free and Community-Driven Critical Infrastructure Ransomware Dataset." In *Proceedings of the International Conference on Cybersecurity, Situational Awareness, and Social Media: Cyber Science 2022; 20–21 June; Wales*, edited by Cyril Onwubiko, Pierangelo Rosati, Anshul Rege, et al. Springer Nature. https://doi.org/10.1007/978-981-19-6414-5_2.
- Romanosky, Sasha. 2016. "Examining the Costs and Causes of Cyber Incidents." *Journal of Cybersecurity* 2 (2): 121–35. <https://doi.org/10.1093/cybsec/tyw001>.
- Romanosky, Sasha, Lillian Ablon, Andreas Kuehn, and Therese Jones. 2019. "Content Analysis of Cyber Insurance Policies: How Do Carriers Price Cyber Risk?" *Journal of Cybersecurity* 5 (1): tyz002. <https://doi.org/10.1093/cybsec/tyz002>.
- Spyromitros-Xioufis, Eleftherios, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. 2012. "Multi-Label Classification Methods for Multi-Target Regression." Preprint, Preprint, arXiv, arXiv:1211.6581.
- Tsohou, Aggeliki, Vasiliki Diamantopoulou, Stefanos Gritzalis, and Costas Lambrinoudakis. 2023. "Cyber Insurance: State of the Art, Trends, and Future Directions." *International Journal of Information Security* 22 (3): 737–48. <https://doi.org/10.1007/s10207-023-00660-8>.
- Tsoumakas, Grigorios, and Ioannis Katakis. 2008. "Multi-Label Classification: An Overview." In *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. IGI Global. <https://doi.org/10.4018/978-1-59904-951-9.ch006>.
- Verizon RISK Team. 2025. "VCDB: VERIS Community Database." <https://github.com/vz-risk/VCDB>.
- Wheatley, Spencer, Annette Hofmann, and Didier Sornette. 2021. "Addressing Insurance of Data Breach Cyber Risks in the Catastrophe Framework." *The Geneva Papers on Risk and Insurance—Issues and Practice* 46 (1): 53–78. <https://doi.org/10.1057/s41288-020-00163-w>.
- Xu, Maochao, Kristin M. Schweitzer, Raymond M. Bateman, and Shouhuai Xu. 2018. "Modeling and Predicting Cyber Hacking Breaches." *IEEE Transactions on Information Forensics and Security* 13 (11): 2856–71. <https://doi.org/10.1109/TIFS.2018.2834227>.
- Zängerle, Daniel, and Dirk Schiereck. 2023. "Modelling and Predicting Enterprise-Level Cyber Risks in the Context of Sparse Data Availability." *The Geneva Papers on Risk and Insurance—Issues and Practice* 48 (2): 434–62. <https://doi.org/10.1057/s41288-022-00282-6>.
- Zhang, Linfeng, Changyue Hu, and Zhiyu Quan. 2025. "NLP-Powered Repository and Search Engine for Academic Papers: A Case Study on Cyber Risk Literature with CyLit." *North American Actuarial Journal* 29 (2): 390–421. <https://doi.org/10.1080/10920277.2024.2416903>.

APPENDICES

APPENDIX A. NOTATION

Table 1. Notation used in multilabel classification and multi-output regression.

Symbol	Description
$\mathcal{D}$	Dataset
$\mathbf{x}_i \in \mathbb{R}^d$	Feature vector for the i -th observation
$\mathbf{X} \in \mathbb{R}^{m \times d}$	Full feature matrix (all m observations)
$\mathbf{X}_{\text{train}} \in \mathbb{R}^{r \times d}$	Training feature matrix
$\mathbf{X}_{\text{test}} \in \mathbb{R}^{s \times d}$	Test feature matrix
$\mathbf{X}_{\text{val}} \in \mathbb{R}^{v \times d}$	Validation feature matrix (split from $\mathbf{X}_{\text{train}}$ via cross-validation)
$\mathbf{y}_i \in \mathcal{Y} = \{0, 1\}^q$	Binary label vector for the i -th observation
$\mathbf{Y}_{\text{train}} \in \{0, 1\}^{r \times q}$	Training label matrix
$\mathbf{Y}_{\text{test}} \in \{0, 1\}^{s \times q}$	Test label matrix
$\mathbf{Y}_{\text{val}} \in \{0, 1\}^{v \times q}$	Corresponding label matrix for the validation fold
$f_{j,\theta}$	Classifier/regressor for label L_j trained with hyperparameter θ
$\mathbf{p}_j^{\text{val}} \in [0, 1]^v$	Predicted probabilities for label j on validation set
$\mathcal{D}_j$	Training dataset for label L_j
$\mathcal{L} = \{L_1, \dots, L_j, \dots, L_q\}$	Label set with q classification output variables
$\hat{\mathbf{y}}_i \in \{0, 1\}^q$	Predicted label vector for the i -th observation
$\mathcal{T}$	Thresholding strategy (FIXED or ADAPTIVE)
$\mathcal{B}$	Base learner family (e.g., random forest)
Θ	Hyperparameter search space
F	Evaluation metric
m	Number of total observations in the dataset
r	Number of training observations
s	Number of test observations
v	Number of validation observations (within training folds)
d	Number of input features
q	Number of labels (output dimensions)
$\mathbf{x}_i^* \in \mathbb{R}^d$	Test observation for the i -th observation
$\hat{\mathbf{y}}_i^* \in \{0, 1\}^q$	Final predicted binary label vector for the i -th observation
$\boldsymbol{\tau}^* = (\tau_1^*, \dots, \tau_q^*)$	Vector of label-specific thresholds
$\hat{\mathbf{Y}}_{\text{test}} \in \{0, 1\}^{s \times q}$	Predicted binary label matrix for the test set
K	Number of cross-validation folds
$\mathbf{z}_i \in \mathbb{R}^q$	Continuous output vector for the i -th observation (e.g., incident counts)
$\mathcal{O} = \{O_1, \dots, O_j, \dots, O_q\}$	Label set with q regression output variables
$\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_q)$	Permutation of $\{1, \dots, q\}$ labels
Σ	Total $q!$ permutations for the labels

APPENDIX B. CLASSIFICATION VALIDATION METRICS

Table 2. Multilabel classification metrics with formulas and notations.

Metric	Description	Formula
Weighted-F1	F1 weighted by label frequency	$\sum_{j=1}^q w_j \cdot F1_j$, where $w_j = \frac{\text{Support}_j}{\sum_k \text{Support}_k}$, $\text{Support}_j = \text{TP}_j + \text{FN}_j$
Macro-F1	Unweighted mean of per-label F1	$\frac{1}{q} \sum_{j=1}^q F1_j$, where $F1_j = \frac{2 \cdot \text{TP}_j}{2 \cdot \text{TP}_j + \text{FP}_j + \text{FN}_j}$
Micro-F1	Global F1 across all labels	$\frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$
Sample-F1	Mean of sample-wise F1 scores	$\frac{1}{m} \sum_{i=1}^m F1_i$, where $F1_i = \frac{2 \cdot \text{TP}_i}{2 \cdot \text{TP}_i + \text{FP}_i + \text{FN}_i}$
Jaccard Index	Intersection over union of label sets	$\frac{1}{m} \sum_{i=1}^m \frac{ \mathcal{S}_i \cap \hat{\mathcal{S}}_i }{ \mathcal{S}_i \cup \hat{\mathcal{S}}_i }$, where $\mathcal{S}_i := \{j : y_{ij} = 1\}$, $\hat{\mathcal{S}}_i := \{j : \hat{y}_{ij} = 1\}$
Hamming Loss	Proportion of incorrect labels	$\frac{1}{mq} \sum_{i=1}^m \sum_{j=1}^q \mathbb{I}(y_{ij} \neq \hat{y}_{ij})$

APPENDIX C. REGRESSION VALIDATION METRICS

Table 3. Multi-output regression metrics with formulas and notations.

Metric	Description	Formula
average MSE (aMSE)	Mean squared error across all outputs and observations	$\text{aMSE} = \frac{1}{q} \sum_{j=1}^q \frac{1}{m} \sum_{i=1}^m (z_{ij} - \hat{z}_{ij})^2$
average RMSE (aRMSE)	Square root of aMSE per output	$\text{aRMSE} = \frac{1}{q} \sum_{j=1}^q \sqrt{\frac{1}{m} \sum_{i=1}^m (z_{ij} - \hat{z}_{ij})^2}$
average Correlation Coefficient (aCC)	Mean Pearson correlation over outputs	$\text{aCC} = \frac{1}{q} \sum_{j=1}^q \frac{\sum_{i=1}^m (z_{ij} - \bar{z}_j)(\hat{z}_{ij} - \bar{\hat{z}}_j)}{\sqrt{\sum_{i=1}^m (z_{ij} - \bar{z}_j)^2} \cdot \sqrt{\sum_{i=1}^m (\hat{z}_{ij} - \bar{\hat{z}}_j)^2}}$
average Relative RMSE (aRRMSE)	RMSE normalized by true std per output	$\text{aRRMSE} = \frac{1}{q} \sum_{j=1}^q \sqrt{\frac{\sum_{i=1}^m (z_{ij} - \hat{z}_{ij})^2}{\sum_{i=1}^m (z_{ij} - \bar{z}_j)^2}}$
Global Euclidean Distance (EU_DIST)	Mean distance between predicted and true vectors	$\frac{1}{m} \sum_{i=1}^m \ \mathbf{z}_i - \hat{\mathbf{z}}_i\ _2 = \frac{1}{m} \sum_{i=1}^m \sqrt{\sum_{j=1}^q (z_{ij} - \hat{z}_{ij})^2}$

APPENDIX D. FEATURES FROM INSURTECH COMPANY

Table 4. Category-level summary of the 527 insurtech features used in analysis.

Category	Explanation	Examples	Number of Features
Business Information	Basic information describing business operations, including address, founding year, opening hours, menu, and other operational features	<i>ADDRESS.coordinates.lat, BUSINESS.founded_year, MENU.has_menu_True</i>	114
Firmographics	Features describing company size, workforce, ownership, revenue and sales, licensing count, and demographic features	<i>FIRMO.company_type_Private, FIRMO.employee_range_11-50, FIRMO.revenue_range_1-4.9Mil, RANDOM_STATS.is_female_owned_False</i>	50
Classification	Categorical segmentation features describing the business segment and segment proportions	<i>CLASSIFICATION.segment_Retail, CLASSIFICATION.segment_Automotive, CLASSIFICATION.segment_details.proportion_segment_Manufacturing</i>	48
Risk Characteristics	Flags representing hazardous or risk-related business activities, services, physical features, and operational exposures	<i>CHARACTERISTICS.business_characteristics.flagged_names.Alcohol, BUSINESS.professional_misconducts_count, URLS.has_non_corporate_email_True</i>	211
Index	Indicators derived from customer reviews that reflect various aspects of business quality, reputation, and service performance	<i>CARPE_INDEXES.customer_rating.score_version_mean, REVIEW_SCORE.average_score, REVIEW_SUMMARY.Overall Business Rating</i>	95
Proximity Score	Environment-based indicators that measure risk arising from nearby businesses and surrounding neighborhood activity	<i>negative_keyword_score.score, proximity_scores.combustibles.density_v1.0, proximity_scores.traffic_mode.distance_v1.0</i>	9

APPENDIX E. CROSS-MODEL FEATURE IMPORTANCE OVERVIEW

Table 5. Summary of the top 10 important features in the classification and regression analyses.

Feature	Cls-Imp	Cls-Const	Reg-Imp	Reg-Const	Feature Category
ACCIDENT_YEAR	✓	✓	✓	✓	Cyber Incident Meta
ADDRESS_coordinates_lat			✓	✓	Business Information
ADDRESS_coordinates_lon			✓	✓	Business Information
ADDRESS_state_east		✓			Business Information
ADDRESS_state_west	✓	✓			Business Information
ADDRESS_street_type_St	✓	✓			Business Information
BUSINESS_founded_year		✓			Business Information
CARPE_SCORES_negative_keyword_score_keyword_count	✓	✓			Proximity Score
CARPE_SCORES_negative_keyword_score_score			✓		Proximity Score
CARPE_SCORES_proximity_scores_traffic_mode_distance_version_1_0			✓	✓	Proximity Score
CHARACTERISTICS_business_characteristics_bc_count	✓	✓	✓	✓	Risk Characteristics
CLASSIFICATION_segment_Finance_Insurance				✓	Classification
CLASSIFICATION_segment_Legal_Services	✓				Classification
CLASSIFICATION_segment_Office_Administrative_Support_Services	✓	✓			Classification
CLASSIFICATION_segment_details_proportion_segment_Finance_Insurance	✓	✓	✓	✓	Classification
CLASSIFICATION_segment_details_proportion_segment_Health_Care_Social_Assistance	✓				Classification
CLASSIFICATION_segment_details_proportion_segment_Legal_Services			✓	✓	Classification
CLASSIFICATION_segment_details_proportion_segment_Office_Administrative_Support_Services	✓	✓	✓	✓	Classification
REVIEW_SCORE_average_score			✓	✓	Index

Notes: (1) ACCIDENT_YEAR is sourced from the cyber incident dataset; all other features originate from the insurtech dataset. (2) *Cls-Imp* indicates whether a feature appears in the top 10 when features are ranked by their *average* log-transformed importance across the different importance measures within each classification model, as summarized in [Figure 3](#). (3) *Cls-Const* captures how consistently a feature is selected across the different importance measures within the same classification model; the top 10 is determined by the rank of counts, corresponding to [Figure 4](#). (4) *Reg-Imp* and *Reg-Const* are defined analogously for regression models, based on the summaries in [Figure 5](#) and [Figure 6](#).

APPENDIX F. UNCONDITIONAL AND CONDITIONAL DEPENDENCE

DEPENDENCE OF OCCURRENCES

The unconditional dependence of occurrences is computed using the occurrence labels in the training set from D2. The conditional dependence of occurrences is computed using the training residuals of LGBM_MOC trained on D2. The association of residuals represents the dependence after the effect of firm-specific characteristics is removed. The phi coefficient is used to measure the degree of pairwise dependence of binary variables.

DEPENDENCE OF FREQUENCIES

The unconditional dependence of frequencies is computed using the frequency labels in the training set from D2. The conditional dependence of frequencies is computed using the training residuals of LGBM_MOR trained on D2. The association of residuals represents the dependence after the effect of firm-specific characteristics is removed. Spearman's rho is used to measure the degree of pairwise dependence of numerical variables.

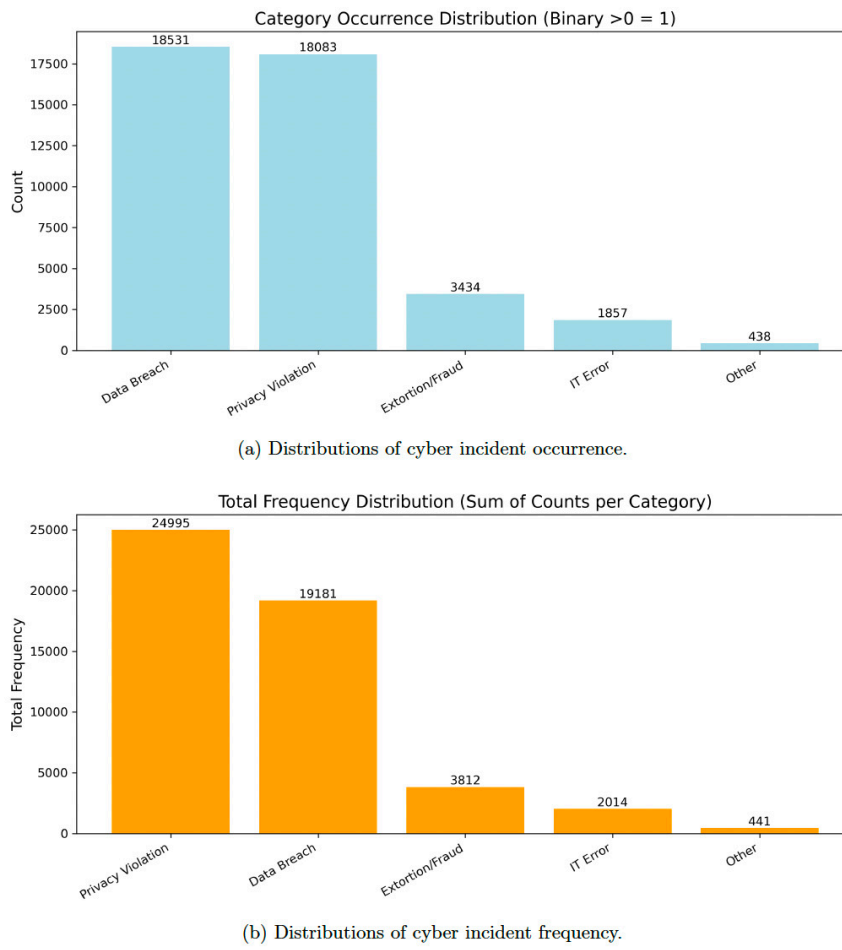
Table 6. Unconditional dependence and conditional dependence of occurrences across incident types.

	Unconditional Dependence of Occurrences (Phi Coefficient)					Conditional Dependence of Occurrences (Phi Coefficient)				
	Privacy Violation	Data Breach	Extortion/Fraud	IT Error	Other	Privacy Violation	Data Breach	Extortion/Fraud	IT Error	Other
Privacy Violation	1.00	-0.75	-0.24	-0.16	-0.07	1.00	-0.02	0.02	0.02	0.02
Data Breach	-0.75	1.00	-0.13	-0.13	-0.04	-0.02	1.00	0.03	0.05	0.03
Extortion/Fraud	-0.24	-0.13	1.00	0.02	-0.02	0.02	0.03	1.00	0.03	0.00
IT Error	-0.16	-0.13	0.02	1.00	-0.01	0.02	0.05	0.03	1.00	-0.01
Other	-0.07	-0.04	-0.02	-0.01	1.00	0.02	0.03	0.00	-0.01	1.00

Table 7. Unconditional dependence and conditional dependence of frequencies across incident types.

	Unconditional Dependence of Frequencies (Spearman's Rho)					Conditional Dependence of Frequencies (Spearman's Rho)				
	Privacy Violation	Data Breach	Extortion/Fraud	IT Error	Other	Privacy Violation	Data Breach	Extortion/Fraud	IT Error	Other
Privacy Violation	1.00	-0.72	-0.23	-0.15	-0.07	1.00	-0.60	-0.09	-0.04	0.01
Data Breach	-0.72	1.00	-0.13	-0.13	-0.04	-0.60	1.00	-0.14	-0.12	-0.03
Extortion/Fraud	-0.23	-0.13	1.00	0.02	-0.02	-0.09	-0.14	1.00	0.08	0.00
IT Error	-0.15	-0.13	0.02	1.00	-0.01	-0.04	-0.12	0.08	1.00	0.06
Other	-0.07	-0.04	-0.02	-0.01	1.00	0.01	-0.03	0.00	0.06	1.00

APPENDIX G. DISTRIBUTION OF CYBER INCIDENTS

**Figure 7. Distributions of cyber incident occurrence and frequency across categories.**