Predicting Cyber Attacks: How to decide where to invest before it's too late!

Written by Cybeta

Predictive Analytics for Cyber Security[*]

December 10, 2020

Abstract

When deciding where to invest in cyber security, company decision-makers naturally wish to maximize return on investment (ROI). This is difficult without a systematic framework for identifying, assessing, and prioritizing cyber risks and remediation efforts. This article describes a solution: Cybeta predictive analytics produce risk metrics that signal about the likelihood of a future cyber attack. These risk metrics allow companies to easily baseline their cyber risk profile against competitors and their industry as a whole, and invest accordingly. In this article, we use Cybeta cyber risk metrics from over 1,700 companies between December 2015 and April 2020, and show that these are predictive of out-of-sample future successful cyber attacks. Our analysis suggests that companies with the highest Cybeta risk scores are much more likely to experience a future successful cyber attack than those with the lowest risk scores. To the best of our knowledge, this is the first study of its kind and the first to document out-of-sample predictability of future cyber attacks. These results highlight the ability of Cybeta risk metrics to overcome many of the limitations inherent in other cyber security risk assessments, allowing companies to target their cyber investments to maximize ROI.

[*] For more information please contact us via email or phone: contact@cybeta.com or 1-779-429-2382.

1. Introduction

The cybersecurity market is forecast to increase from \$119.9 billion in 2019 to \$433.6 billion by 2030, at a 12.6% CAGR between 2020 and 2030.[1] Despite this vast expenditure, companies continue to experience breaches every day. The failure to prevent such successful cyber attacks lies in the inability of organizations to anticipate what is likely to occur, and for what reason. The most common approach used to try to identify cyber risks involves scanning a network to identify vulnerabilities or indicators of compromise from past breaches, overlaid with various security controls. This methodology, while intuitive, fails to prevent a future cyber breach because it only takes into account what is happening right now. Cyber attackers are continuously adapting and an effective cyber security approach must match the adaptive nature of the threat.

When making a cyber security investment, the ultimate decision should be a result of cost-benefit analysis. The challenge with conducting such an analysis in cyber security is that measuring potential benefits is non-trivial. The standard approach on its own does not provide the necessary information to conduct a cost-benefit analysis and identify which vulnerabilities should be addressed first. Cybeta solves this challenge by providing a continuously adaptive cyber-focused, data driven framework for projecting future breach likelihood.

Consider a simplified model of how to determine a type of return on investment (ROI) for a cyber security investment. Let V be the cost of the investment (the spend required to alleviate vulnerabilities) and let C be the expected cost to the company, if a successful cyber attack were to occur. Finally, let p and $p(V)$ be the probability of a successful cyber attack before and after the investment respectively. The ROI then has a straightforward form: $ROI = \frac{(p-p(V))\times C}{V}$.[2] Clearly, the probability of an attack, p,

[1] https://www.globenewswire.com/news-release/2020/11/05/2120926/0/en/The-cyber-security-market-will-grow-from-119-9-billion-in-2019-to-433-6-billion-by-2030-at-a-12-6-CAGR-between-2020-and-2030.html

[2] This formula is a byproduct of the following. Consider the case when the company does not make an investment, in this case the expected cost of the breach is $p \times C$. That is, the expected cost is the cost of the breach, C, times the probability, p, of the breach occurring. When the company does make the

and the potential reduction from making the investment, $p(V)$, are critical in evaluating the ROI of a cyber investment. In fact, without these probabilities, it is impossible to calculate an ROI. Using the traditional approach of assessing cyber risk, it might be possible to obtain some information about V and C. However, the probability of an attack is impossible to predict from a simple list of vulnerabilities.

Cybeta takes advantage of advanced algorithms that provide signals about the probability of a successful future cyber attack. These unique algorithms have been developed by experts with extensive industry expertise and with analytics to provide signals about future cyber security breaches. This article provides rigorous tests, using out-of-sample data, of the predictive power of these signals.

Our main findings are that companies with the highest Cybeta risk scores (i.e., those companies that are at the highest risk of a cyber breach) are much more likely to experience a future successful cyber attack than companies with the lowest risk scores. These results are robust to different empirical specifications and are based on a large sample of firms over almost five years.

To the best of our knowledge, this is the first study of its kind and the first to document out-of-sample predictability of future cyber attacks.

2. Cyber Risk Metrics and Predictability of Future Cyber Breaches

Cybeta's main cyber risk metric is Threat Beta$^{\text{TM}}$, which is an indicator of the likelihood of a cyber attack. Threat Beta is based on a proprietary algorithm derived from leading industry expertise and several different sources of cyber security data. Because Threat Beta is argued to be an indicator of a cyber attack, one would expect firms with high Threat Betas to have higher incidences of future cyber breaches than firms with low Threat Betas. That is, Threat Beta risk metrics should be predictive of successful future cyber attacks.

investment, the expected cost of the breach is $p(V) \times C$. Thus, the reduction in expected costs (which is the same as the expected gain from investing in cyber security) is $(p - p(V)) \times C$. Dividing the gain from investment by the spend on investment, V, gives $(p - p(V)) \times C/V$, which represents the ROI.

Before discussing our data and results, the notion of prediction (or predictability) warrants some discussion. When we are testing for predictability, we are not testing for a perfect forecast of the future–we do not mean that we can exactly predict every cyber attack that will take place; this is not possible. Instead, we are testing for a more accurate forecast of the future, on average.

As an analogy, suppose we have two blackjack players. Player P1, knows the rules of the game, and knows the general rules of thumb (i.e., always hit if you have a hard 11 or less, etc.) and player P2, who knows all the same rules, but also knows how to count cards. Because P2 has an advantage, we would expect P2 to win more often than P1, on average. This does not mean that P1 could not win some of the time. The ability to count cards does not guarantee a win every hand – but it does tilt the odds in P2's favor. So if we were to bet on which player, P1 or P2, would win most often, we would bet on P2 and with enough time, this should be confirmed by the data.

Now suppose that we had a third player, P3, who knows nothing about blackjack, is prone to foolish mistakes like hitting on a "hard 20" hand, and does not have the cognitive ability to learn the rules over time. If we knew all of this, then we would be able to rank each of the players we thought would win most often. Specifically, we would say that we expected P3 to win the least often, P1 the second least often, and P2 the most often.

Our predictability tests share a similarity to the above blackjack example, because we are using Cybeta's risk metrics to rank firms based on their cyber risk. If Threat Beta is a predictor of future cyber breaches, then firms with higher Threat Beta's should be associated with a high number of future cyber breaches.

We test for this in two ways: 1) parametric tests using predictive regressions where we regress future cyber breaches on current Threat Betas and test for a statistically significant predictive relation; and 2) non-parametric tests where we sort firms from low to high risk according to current Threat Beta risk scores and test if high Threat Beta firms have higher on average incidences of future cyber breaches than do low Threat Beta firms.

3. Data

For this study we have collected monthly Threat Beta's from Cybeta for 1,716 publicly traded companies from December 2015 through April 2020. These firm-month observations are then matched to reported future cyber breaches of sufficient magnitude to warrant an insurance claim.[3] We end our sample period in April 2020 since this allows us to examine a relation with breaches that occur up to six months after that date (up to October 2020).

We also collect a firm-month "size" variable, market capitalization, which is end-of-month price per share times number of shares outstanding. This size variable, along with industry membership allow us to formulate our tests to ensure that empirical relations seen with future cyber breaches and Cybeta cyber risk metrics are not attributable to either of these important variables.

Table 1 provides summary statistics for the key variables used in the study. It provides the number of observations, mean and standard deviation over time, and overall for both firm Threat Beta's and breaches in our sample and (Panel A) by size (market capitalization) (Panel B). Breaches take the value of a 1 and non-breaches take the value of a 0.

[3] Examples of these cyber breaches include: cyber extortion, malicious data breaches, unauthorized data collection, and fraudulent use/account access.

Table 1: Summary Statistics

Table 1 presents summary statistics of key variables used in the study. CE_{t+1} represents a reported successful cyber breach in month $t+1$ and TB_t represents Cybeta's Threat Beta measured in month t.

(a) Panel A: Summary Statistics Over Time

	N Months	N Obs	Mean CE_{t+1}	Std. CE_{t+1}	Mean TB_t	Std. TB_t
2015	1	1,434	0.029	0.169	1.208	0.429
2016	12	17,479	0.025	0.156	1.226	0.435
2017	12	17,335	0.022	0.147	1.257	0.433
2018	12	17,304	0.024	0.152	1.252	0.434
2019	12	17,234	0.014	0.117	1.262	0.427
2020	4	5,699	0.003	0.056	1.269	0.425
Total	53	76,485	0.020	0.140	1.250	0.432

(b) Panel B: Summary Statistics by Market Capitalization

	N Obs	Mean CE_{t+1}	Std. CE_{t+1}	Mean TB_t	Std. TB_t
Tiny	19,694	0.009	0.092	1.078	0.377
Small	13,873	0.011	0.104	1.172	0.404
Mid	13,606	0.015	0.120	1.201	0.394
Large	13,548	0.018	0.132	1.336	0.420
Mega	15,764	0.049	0.216	1.502	0.429

Overall, Table 1 Panel A shows that of the 76,485 firm-months in our sample, approximately 2.0% experience a future cyber breach. There is a general downward trend in the number of breaches observed over time. This is in part driven by the time lag in the reporting of cyber breach driven insurance lawsuit claims.[4] Threat Beta has a reasonably stable cross-sectional average and standard deviation across the sample period.

Table 1 Panel B shows a clear relation between firm market capitalization (size) and both future cyber breaches and Threat Beta. The largest firms in our sample are more than five times more likely to experience a cyber breach than the smallest firms. The same general trend exists between Threat Beta and firm size, with larger firms having higher cyber risk.

[4] This feature in the data does not work in favor of finding a relation between Threat Beta and future cyber breaches. If anything, it works against us finding a relation.

4. Empirical Tests

4.1. Predictive Regressions

Our first set of tests use predictive regressions where future cyber breaches are regressed on current Threat Beta scores. The regression model we use if of the form:

$$CE_{t+\tau} = \alpha + \beta \times TB_t + \epsilon_{t+\tau}, \tag{1}$$

where $CE_{t+\tau}$ is 1 if a cyber breach occurs within τ months ahead of time t and zero otherwise. TB_t represents a company's current (time t) Threat Beta. In these tests, the key variable of interest is the β parameter. A positive β indicates that firms with a higher current Threat Beta are more likely to experience a future cyber breach.

4.2. Non-Parametric Predictive Tests

Our second set of tests are "non-parametric" tests where we sort firms from low to high Threat Beta and test whether high Threat Beta firms have higher on average incidences of cyber breaches. These tests allow us to observe patterns in the data that are not possible using regression tests and they also serve as an alternative test that can be interpreted as validation of the regression results.

4.3. Results of Predictive Regressions

Table 2 provides the results of our tests using predictive regressions under several different specifications. The rows, e.g., $\tau = 1$, represent tests over different horizons into the future. Where $\tau = 1$ represents one-month-ahead and $\tau = 6$ represents six-months-ahead.[5] The "Estimate" ("Test statistic") row for each horizon is the estimated (level of statistical significance in parenthesis) relation between Threat Beta and future cyber breaches. A Test statistic of greater than 2 is usually regarded as highly significant.

[5] Our reported tests are based on cumulative future breaches, that is $C_{t+\tau} = 1$ if there is at least one breach at any point between t and τ. Our results are qualitatively similar if we use non-cumulative future breaches.

Column (1) shows the main results of the study. The estimated β coefficient is positive and highly statistically significant indicating a strong predictive relation between future successful cyber attacks and current Threat Betas. Importantly the coefficient is positive and significant from one- to six-months into the future, providing strong data driven evidence that Threat Beta is a significant predictor of future cyber breaches up to at least six-months into the future. Columns (2) through (4) provides the results when controlling respectively for date, date and industry, and date, industry, and size. As the table shows, the predictability of Threat Beta is not driven by these important variables, as the β coefficient remains positive and highly statistically significant in all specifications.

Table 2: Predictive Regression Tests

This table reports the estimated relation between current Cybeta risk scores and future cyber breaches ($CE_{t+\tau}$). The logistic regression model is of the form $CE_{t+\tau} = \alpha + \beta \times TB_t + \epsilon_{t+\tau}$. Coefficients, i.e., the estimates of β, are estimated using a logistic regression. Test statistics are in parentheses. Significance levels of 1%, 5%, and 10% are denoted by, ***, **, and *, respectively.

		Regression Model: $CE_{t+\tau} = \beta \times TB_t + \epsilon_{t+\tau}$			
		(1)	(2)	(3)	(4)
$\tau = 1$	Estimate	1.026***	1.046***	1.155***	0.524***
	Test statistic	(18.086)	(18.371)	(19.899)	(8.222)
$\tau = 2$	Estimate	1.026***	1.046***	1.155***	0.524***
	Test statistic	(18.086)	(18.371)	(19.899)	(8.222)
$\tau = 3$	Estimate	0.972***	0.995***	1.103***	0.513***
	Test statistic	(22.176)	(22.596)	(24.480)	(10.398)
$\tau = 4$	Estimate	0.942***	0.966***	1.073***	0.513***
	Test statistic	(24.673)	(25.194)	(27.263)	(11.938)
$\tau = 5$	Estimate	0.925***	0.950***	1.054***	0.514***
	Test statistic	(26.546)	(27.121)	(29.271)	(13.112)
$\tau = 6$	Estimate	0.912***	0.937***	1.038***	0.515***
	Test statistic	(28.023)	(28.668)	(30.850)	(14.074)
Controls:		None	Date	Date&Ind.	Date&Ind.&Size

These represent a formal statistical test of predictability under different specifications and hence are a critical step in the validation of predictability. However, they are limited in their ability provide insight into the degree of predictability, i.e., it is not possible to say that "a one percent increase in Threat Beta implies some percentage increase in the likelihood of a future cyber breach." All we can say is that as Threat Beta increases there is on average a higher frequency of future cyber breaches. Our next set of tests allow us to better quantify the degree of predictability associated with TB and future cyber breaches.

4.4. Results of Non-Parametric Tests

Table 3 provides the results of our non-parametric tests. To conduct these tests, at the end of each month we sort firms based on their Threat Beta score and generate 5 groups from lowest to highest each month. The values reported in the table represent the average proportion (frequency) of breaches within each group over our sample period.

To be most useful as a predictive tool, firms with the lowest Threat Beta scores would experience the lowest number of future cyber breaches and firms with the highest Threat Beta scores would experience the highest number of future cyber breaches. As Table 3 shows, this is exactly what we find. Table 3 column (1) presents the frequency (in percent) of future cyber breaches for firms with the lowest Threat Beta risk scores. Column (5) presents the frequency (in percent) of future cyber breaches for firms with the highest Threat Beta risk scores. There is a clear monotonic relation between Threat Beta and future cyber breaches, across all horizons from one-month-ahead ($\tau = 1$) to six-months-ahead ($\tau = 6$). In all cases, firms with the highest Threat Beta scores have a much higher frequency of cyber breaches than firms with the lowest risk scores.

Column (6) of Table 3 tests formally for a difference between firms in the highest and lowest TB groupings. Focusing on six-months-ahead, firms with very high Threat Betas are more than three times as likely to experience a future cyber breach than firms with very low Threat Betas and this difference is highly statistically significant.

Table 3: Non-Parametric Tests

This table reports the relation between current Cybeta Threat Betas (TBs) and future cyber breaches ($CE_{t+\tau}$). Frequency of future cyber breaches are in percent. Significance levels of 1%, 5%, and 10% are denoted by, ***, **, and *, respectively.

		Threat Beta (TB) →					
		Very Low	Low	Medium	High	Very High	Very High-Very Low
		(1)	(2)	(3)	(4)	(5)	(6)
$\leftarrow CE_{t+\tau} \rightarrow$	$\tau = 1$	1.00%	1.42%	2.12%	3.07%	4.01%	3.00%***
	$\tau = 2$	1.88%	2.57%	3.76%	5.28%	7.04%	5.16%***
	$\tau = 3$	2.68%	3.60%	5.17%	7.13%	9.84%	7.16%***
	$\tau = 4$	3.51%	4.52%	6.40%	8.82%	12.49%	8.98%***
	$\tau = 5$	4.27%	5.47%	7.54%	10.44%	15.08%	10.81%***
	$\tau = 6$	5.01%	6.41%	8.64%	11.93%	17.71%	12.70%***

5. Conclusion

To our knowledge, this is the first study of out-of-sample predictability of successful cyber breaches. The results indicate that Cybeta's quantitative cyber risk metric, Threat Beta, has significant predictive power for future cyber breaches. Firms with high Threat Betas have a much higher probability of a successful future cyber breach than firms with low Threat Betas. These results have direct implications for cyber security investment decisions as Threat Beta offers a validated tool for narrowing in on key vulnerabilities that reduce the probability of a successful cyber attack and maximize the ROI of the investment.