Survival Analysis on US Company Bankruptcy Prediction Dataset

Accounting Data of NYSE and NASDAQ Companies (1999-2018)

Hao Li, Haonan Gu, Jingxi Li, Lijie He
Columbia University, 2025
May 13, 2025

Report submitted for STATS 5231 SURVIVAL ANALYSIS

1 Introduction

Corporate bankruptcy prediction is a critical area in financial risk management with significant implications for investors, creditors, and policymakers. In this study, we apply survival analysis techniques to examine the time until bankruptcy for U.S. companies, based on the American Companies Bankruptcy Prediction Dataset from Kaggle [5]. The dataset comprises 78,682 firm-year observations for 8,971 unique companies observed annually between 1999 and 2018. Each observation contains 18 financial ratios capturing key aspects of liquidity, profitability, leverage, and operational efficiency. To see the meaning of the variables in the data, please read the table in the Attachment Section.

Of the total observations, 5,220 (6.6%) correspond to bankruptcy events, while the remaining 73,462 (93.4%) are treated as censored at that time point. A preliminary examination of the data reveals substantial multicollinearity among financial predictors and a non-exponential Kaplan-Meier survival curve, suggesting the need for appropriate methodological adaptations.

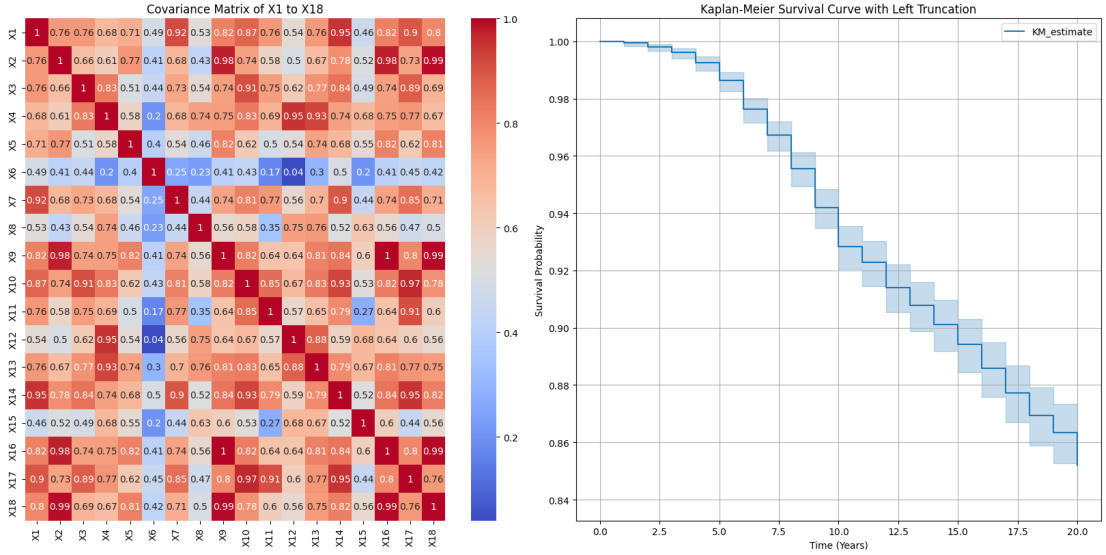


Figure 1: Data Overview: Covariates with co-linearity (left) and Kaplan-Meier survival curve with censoring (right)

Our research addresses three primary questions: (1) **Memorylessness**: Is corporate bankruptcy a memoryless process, or does it depend on the company's age and financial history? (2) **Risk factors**: Which financial indicators most strongly predict corporate insolvency, and how do these risks evolve over time? (3) **Predictive accuracy**: To what extent can we accurately forecast bankruptcy within fixed time horizons using historical financial data?

2 Methods

2.1 Time-to-Event Analysis Framework

The event of interest is a company's failure (bankruptcy), with observations potentially left-truncated (companies not observed since 1999) and right-censored (companies remaining solvent at the end of 2018). We analyze this data using both parametric and semi-parametric survival models, with careful consideration of censoring mechanisms.

An important methodological consideration is that we focus on using financial metrics at the initial observation time to predict subsequent bankruptcy risk. This approach prevents data leakage from future observations and better reflects real-world prediction scenarios where future financial states are unknown.

2.2 Distributional Testing and Parametric Models

We first test whether company failures follow an exponential distribution by comparing it against the more flexible Weibull distribution. The exponential distribution would indicate memorylessness (constant hazard rate), whereas the Weibull allows for time-dependent hazard rates. The memorylessness property is characterized by: $P(T > s + t|T > s) = P(T > t)$

We also compare against the log-normal distribution using likelihood-based measures (AIC). These models provide a foundation for understanding the baseline hazard function before incorporating covariates. Implementation is based on Therneau's survival package [4] and methods described in Therneau and Grambsch [3].

2.3 Cox Proportional Hazards Model

We implement a time-invariant effect linear Cox model of the form: $\lambda(t|X) = \lambda_0(t)\exp(\beta^T X)$ where $\lambda_0(t)$ is the baseline hazard function, X represents the vector of financial covariates, and β contains the corresponding coefficients.

To address the multicollinearity among financial predictors, we employ several feature selection approaches: (1) Recursive Feature Elimination based on AIC (AIC $= 2k - 2\ell(\hat{\theta})$) (2) L1/L2 penalization with the objective function: $-\ell(\beta) + \lambda(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2)$ (3) Exhaustive search comparing all possible combinations of the 18 covariates

We also conduct proportional hazards tests to verify the key assumption of the Cox model.

2.4 Machine Learning Approaches

For non-linear modeling, we implement two advanced approaches:

Survival Forests: Using the LTRC forest package [2], which accommodates left-truncated and right-censored data and recursively partitions the covariate space into regions with distinctive hazard rates.

Neural Networks: We explore a Transformer-based neural network model designed to process variable-length historical sequences of financial ratios and predict failure within a two-year window. This approach explicitly incorporates the temporal dynamics of financial indicators, unlike the static Cox model.

For prediction evaluation, we implement a temporal validation scheme, fitting models on data before a cutoff year (2013) and evaluating performance on subsequent observations. To address the class imbalance problem, we examine both the original and class-balanced datasets.

3 Results

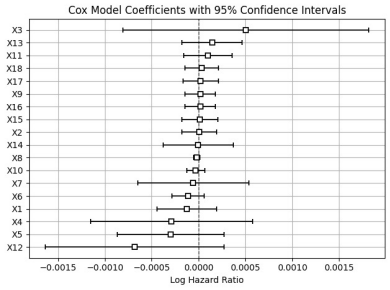
3.1 Distributional Analysis

The exponential distribution is strongly rejected in favor of the Weibull alternative (p-value: 4.69×10^{-37}), providing clear evidence that company survival is not memoryless. The log-normal distribution provides a slightly better fit than the Weibull model (Log-Normal AIC: 6960.39 vs. Weibull AIC: 6994.31), suggesting that bankruptcy hazard is initially increasing and then decreasing over time.

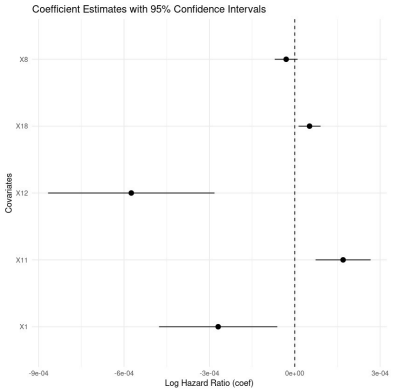
3.2 Cox Model Feature Selection

A Cox model with all 18 covariates identifies no significant predictors. This is normal because great number of features will cause insignificance in colinearity situations. So we want to reduce factors by feature selection. Traditional forward and backward feature selection methods (using the MASS package [1]) proved ineffective at eliminating redundant variables, likely due to the high correlation structure as you could see in *stepwise_cox, ipynb*. So we did an exhaustive search identified optimal

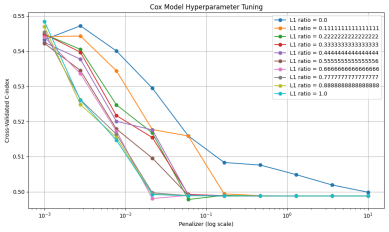
combinations of features based on AIC (X1+X8+X11+X12+X18), with confidence intervals shown in Figure 2. The Cox proportional hazards model identifies three statistically significant predictors of company bankruptcy risk: X12 significantly decreases bankruptcy risk (negative coefficient with confidence interval below zero), while X11 and X18 significantly increase bankruptcy risk (positive coefficients with confidence intervals above zero), with X1 showing a borderline protective effect and X8 having minimal impact on survival.



(a) Complete Cox model coefficients with all predictors



(b) Confidence intervals of coefficients for features selected in Cox models



(c) Effect of L1, L2 penalization on C-index

Figure 2: Cox model selection and optimization results

The Cox model hyperparameter tuning results show that predictive performance (measured by concordance index) consistently decreases as regularization strength increases across all L1 ratio values, with the best performance ($\sim$ 0.545 C-index) achieved using minimal regularization, suggesting that while regularization helps prevent overfitting, excessive penalization significantly reduces the model's ability to discriminate between companies that will and won't experience bankruptcy.

Also, after we deal with the high colinearity (see attachments for detail), the proportional hazards assumption was not violated, indicating that the selected financial ratios have a consistent effect on survival over time.

3.3 Machine Learning Models

The survival forest approach showed moderate predictive power with an ROC area under the curve of approximately 0.62. A key limitation is that this model, like the Cox model, only utilizes financial ratios at entry time rather than their evolving values.

Our Transformer-based neural network, which incorporates time-series financial data, demonstrated strong predictive performance on the balanced dataset (1:1 failure-to-survival ratio). However, when applied to the original imbalanced dataset, the model tended to over-predict failures. This suggests that while sequential modeling improves bankruptcy prediction, substantial challenges remain in calibrating predictions for rare events.

The neural network achieved an F1 score of approximately 0.75 on the balanced test set but dropped to 0.3 on the original imbalanced dataset. This performance gap highlights the inherent difficulty in predicting rare bankruptcy events, even with sophisticated modeling approaches.

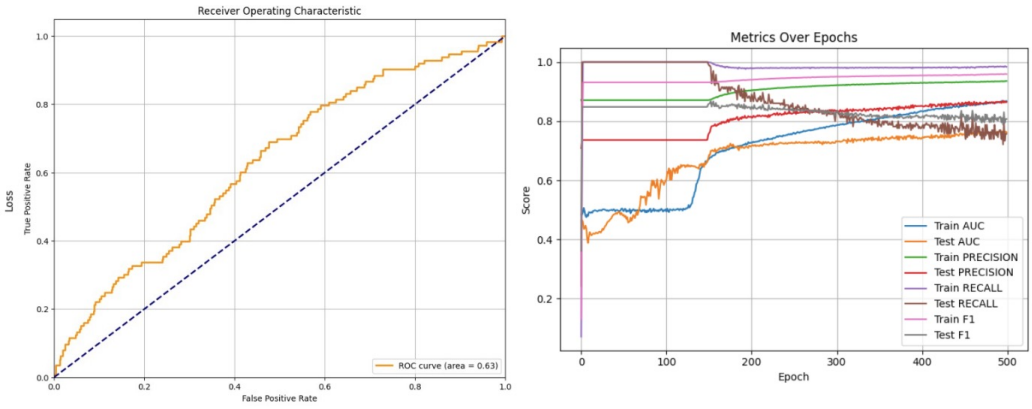


Figure 3: Left: ROC of survival forest, Right: Metrics of neural network model on conditional distribution

4 Conclusions

Our analysis reveals several important insights into corporate bankruptcy prediction:

Non-memoryless process: While all 18 covariates have consistent effects on bankruptcy effects over time, company survival itself is definitively not a memoryless process, with bankruptcy hazard dependent on company age and evolving over time.

Key financial predictors: Among 18 financial ratios, the linear combination of (X1+X8+X11+X12+X18) is the most significant predictors of bankruptcy.

Prediction limitations: While our models achieve respectable predictive performance metrics, particularly in balanced datasets, their practical utility remains limited by challenges in calibrating predictions for rare events. The tendency to over-predict failures in realistic, imbalanced settings indicates inherent limitations in bankruptcy forecasting from public financial data.

Market efficiency implications: The difficulty in achieving high-precision bankruptcy prediction aligns with efficient market hypotheses. If bankruptcies were easily predictable from public data, such signals would be quickly arbitraged away through financial instruments like put options or would enable companies to take preemptive actions, changing the outcome.

These findings have important implications for financial risk assessment, suggesting that while certain combination of covariates serve as meaningful indicators of corporate financial health, precise bankruptcy timing remains challenging to predict. Future research could explore incorporating macroeconomic factors, industry-specific indicators, and alternative data sources to enhance predictive models.

References

[1] Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

[2] Yao, W., Frydman, H., Lafosse-Marin, D., & Simonoff, J. S. (2022). Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research*, 31(11), 2217-2236.

[3] Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer. ISBN 0-387-98784-3.

[4] Therneau, T. M. (2024). A Package for Survival Analysis in R. R package version 3.8-3.

[5] Utkarshx27. (2023). American Companies Bankruptcy Prediction Dataset. Kaggle. `https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset`.

5 Attachments

See Zip File attached on Courseworks for more details.

Code to select features exhaustively, or randomly is separately running in 'search_factors.r' and 'exhausive.r'.

Pictures of covariance matrix, Kaplan-Meier Survival Curve, Cox model coefficients, Cox Model Hyperparameter Tuning can be found in 'SurvivalFinal.ipynb'.

Proportional Hazard Ratio Test and ROC Curve of random survival forest can be found in randomsurvivalforest.ipynb.

Metrics over Epochs can be found in advanced_predictors.ipynb.

Confidence intervals of coefficients for features selected in Cox models can be found in stepwise_cox.ipynb.

The three csv files are a result of slightly different processing of data using different methods of four authors.

To see figures, outputs, and results, you can click on the Folder "figures and outputs".

Table 1: Description of Variables in American Bankruptcy Dataset

Variable	Type	Description
Response Variable		
status_label	Categorical	Indicator of company status: 'failed' for bankrupt companies; otherwise active/solvent. Used to create the binary 'status' variable (1=failed, 0=active) for survival analysis.
Predictor Variables (Financial Ratios)		
X1	Numerical	Total Assets - Represents the size of the company and its total resources.
X2	Numerical	Liquidity ratio - Measures company's ability to pay short-term obligations.
X3	Numerical	Depreciation/Amortization - Reflects non-cash expenses related to assets.
X4	Numerical	EBITDA (Earnings Before Interest, Taxes, Depreciation & Amortization) - Indicates operational profitability.
X5	Numerical	Leverage ratio - Measures the extent of company's financing by debt.
X6	Numerical	Activity ratio - Indicates efficiency of asset utilization.
X7	Numerical	Cash flow indicator - Measures cash generated from operations.
X8	Numerical	Profitability ratio - Indicates return on assets or equity.
X9	Numerical	Growth rate - Measures year-over-year expansion of revenues or assets.
X10	Numerical	Capital structure ratio - Indicates proportion of different funding sources.
X11	Numerical	Cash Flow to Debt ratio - Indicates ability to pay debt with operating cash flow.
X12	Numerical	Return on Investment ratio - Indicates efficiency of invested capital.
X13	Numerical	Working capital ratio - Measures operational liquidity.
X14	Numerical	Asset turnover ratio - Measures efficiency of assets in generating revenue.
X15	Numerical	Interest coverage ratio - Measures ability to pay interest expenses.
X16	Numerical	Revenue/Sales metric - Indicates top-line performance.
X17	Numerical	Debt service coverage ratio - Indicates ability to meet debt obligations.
X18	Numerical	Market-to-book ratio - Compares market value to book value.
Time-Related Variables		
year	Numerical	Calendar year of observation (1999-2018).
survival_time	Numerical	Number of years from 1999 until bankruptcy or censoring.
entry_time	Numerical	Number of years from 1999 until company enters study.