A Survival Analysis of Determinants for Startup Failures in China

Zhejun Xiao
Institute of Statistics and Big Data
2025104263@ruc.edu.cn

December 27, 2025

Abstract

This report analyzes the lifecycle of 6,271 failed companies to understand the determinants of startup survival within the Chinese ecosystem. We employ a multi-methodological approach combining **Cox Proportional Hazards (CPH) Model** for risk factor analysis and **Machine Learning (XGBoost-/Random Forest)** for lifespan prediction. Our Cox model, validated by a C-index of 0.8938, reveals that financial capability is the strongest predictor of longevity, though its protective effect diminishes over time (Schoenfeld residual violation). Furthermore, **XGBoost outperforms Random Forest** in predicting the exact survival days (RMSE=342.5), demonstrating superior capability in capturing non-linear interactions between industry features and funding rounds.

1 Introduction and Data Overview

1.1 Problem Statement and Analytical Goal

Understanding why and, critically, **when** startups fail is a key challenge for investors, policymakers, and entrepreneurs. The lifecycle of a startup is often characterized by a high-risk "Valley of Death" phase. This project utilizes a rich dataset of 6,271 failed Chinese companies. Our primary goal is two-fold:

1. Apply **Survival Analysis** to quantify the instantaneous risk of failure and identify statistically significant risk factors.
2. Implement **Machine Learning Algorithms** to build a predictive model for the exact lifespan (days) of a startup, comparing the efficacy of ensemble methods.

1.2 Data Preprocessing and Feature Engineering

To ensure code robustness and modularity, the analysis framework was encapsulated within a custom Python class `StartupAnalysisSystem`. The preprocessing pipeline included:

1. **Cleaning:** Removed 5 empty columns and filtered out anomalies where `live_days` ≤ 0, resulting in a valid sample size of 6,271.
2. **Imputation:** Missing values in `death_reason` were marked as "Unknown" to maintain the sample size for subsequent text-mining and cause-specific analysis.
3. **Feature Engineering and Grouping:** Categorical variables (Industry, Location) had numerous sparse categories. To ensure the robustness of the CPH model, we grouped the top 5 most frequent categories into their respective groups and collapsed the rest into an 'Other' group. Funding rounds were mapped to an ordinal scale (`financing_level`), and a binary flag (`has_funding`) was created.

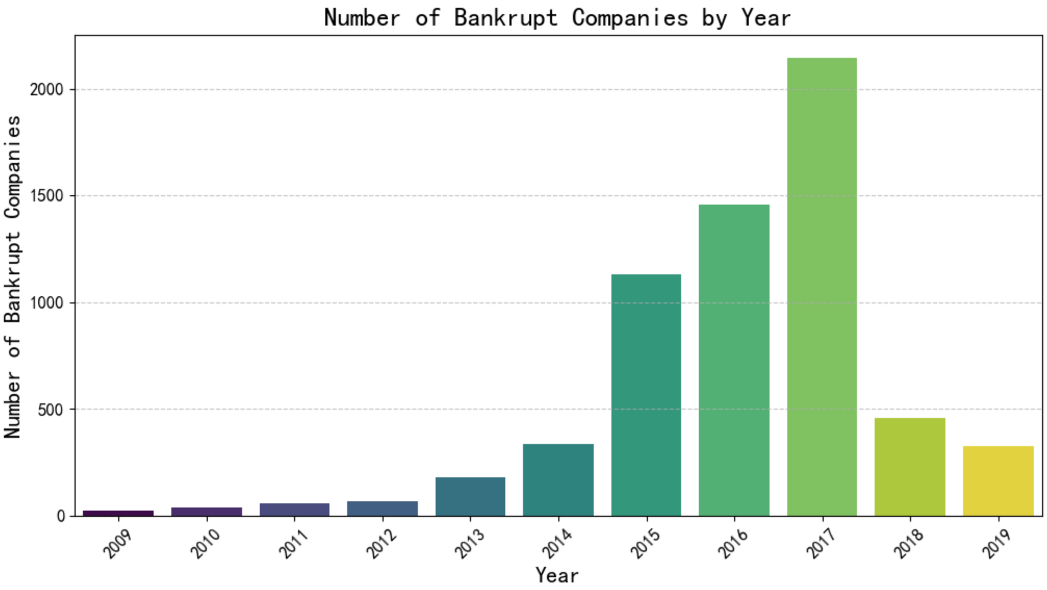


Figure 1: Number of Closed Companies by Year.

1.3 Macro-Level Trends

The Exploratory Data Analysis (EDA) reveals several macro trends reflecting the underlying market dynamics:

- **Temporal Trend (Figure 1):** The failure distribution peaked around 2017. This surge lags 2-3 years behind the 2014-2015 "Mass Entrepreneurship" founding boom, suggesting that the typical startup survival bottleneckwhere early capital runs out or a sustainable business model fails to materializemanifests around the 3-year mark.

- **Geographic Concentration (Figure 2):** Failures are highly concentrated in first-tier cities (Beijing, Guangdong, Shanghai). This concentration reflects not only where startup activity is highest but also the regions with the most intense competition, high operational costs, and rapid market velocity.

- **Industry and Cause (Figures 3 & 4):** E-commerce and Enterprise Services lead in failure counts. Critically, "Business Model Failure" far outstrips "Cash Flow Break," supporting the view that while lack of money is the immediate cause of death, fundamental flaws in the value proposition or monetization strategy (business model) are the root systemic issues.

2 Exploratory Analysis of Survival Lifecycle

2.1 The Startup "Valley of Death"

Visualizing the distribution of survival days (Figure 5) confirms the characteristic right-skewness of survival data. The median survival time is 1,121 days (approximately 3 years and 2 months). The high density of failures occurring before this median highlights the existence of a severe "Valley of Death," where companies that survive past the 3-year mark are likely to have established some level of product-market fit or secured further funding, significantly improving their odds of long-term survival.

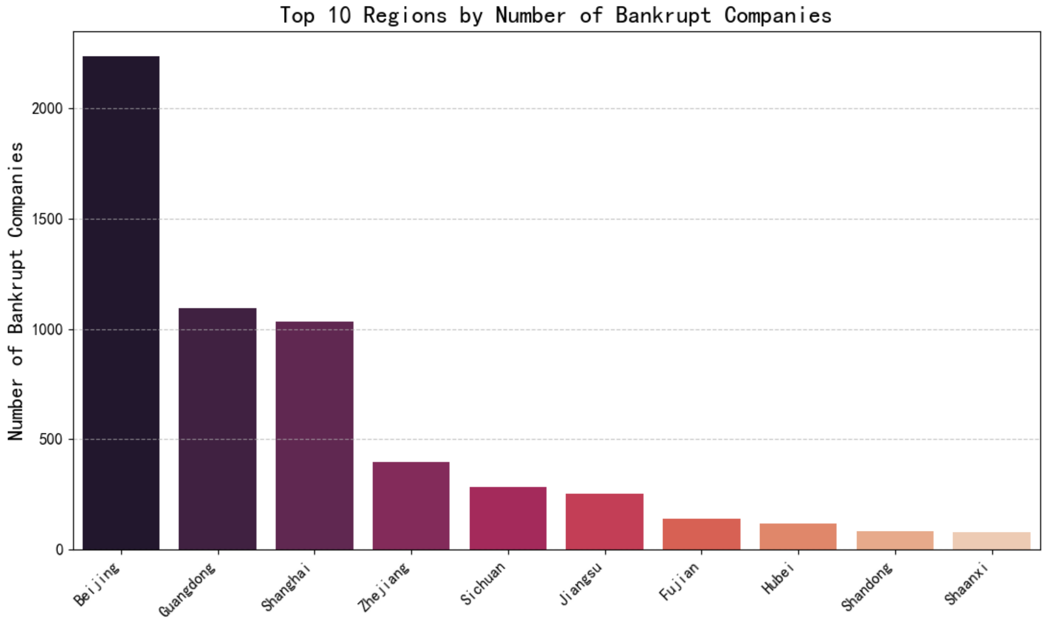


Figure 2: Top 10 Regions by Number of Failures. Beijing, Guangdong, and Shanghai dominate the list.

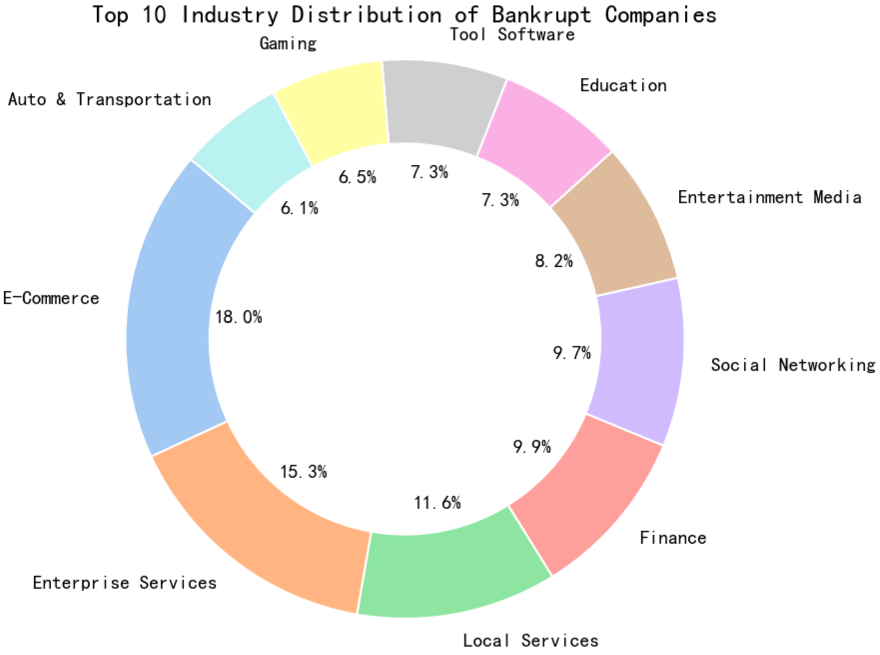


Figure 3: Top 10 Industries Distribution.

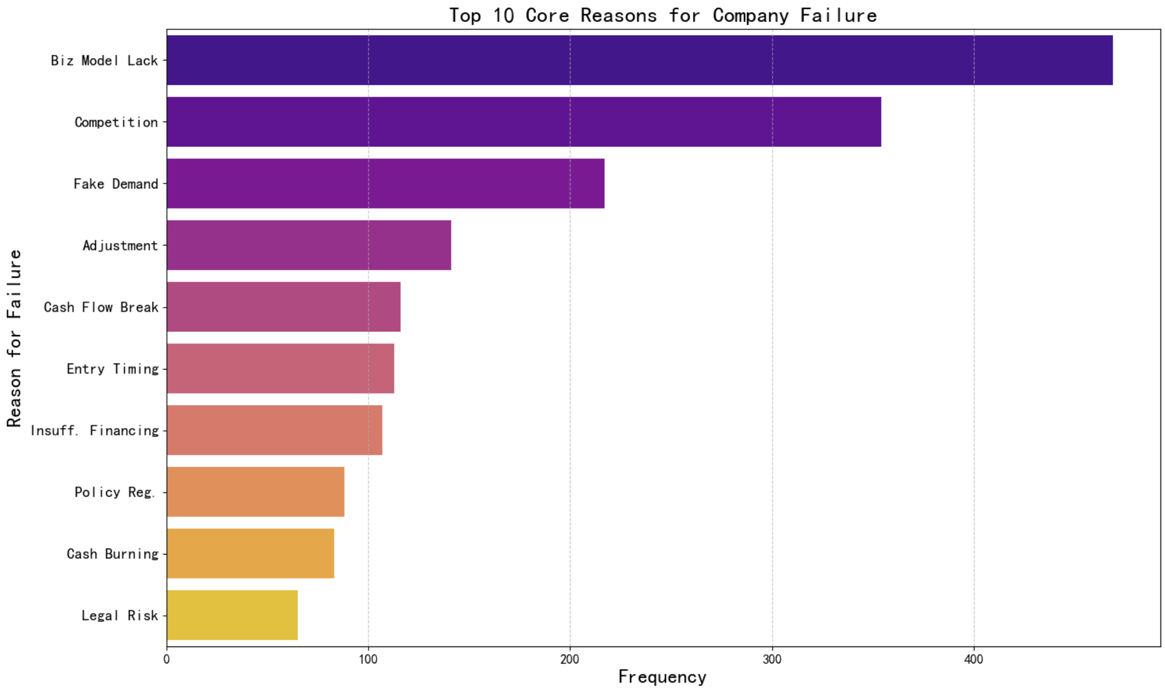


Figure 4: Top 10 Core Reasons for Failure.

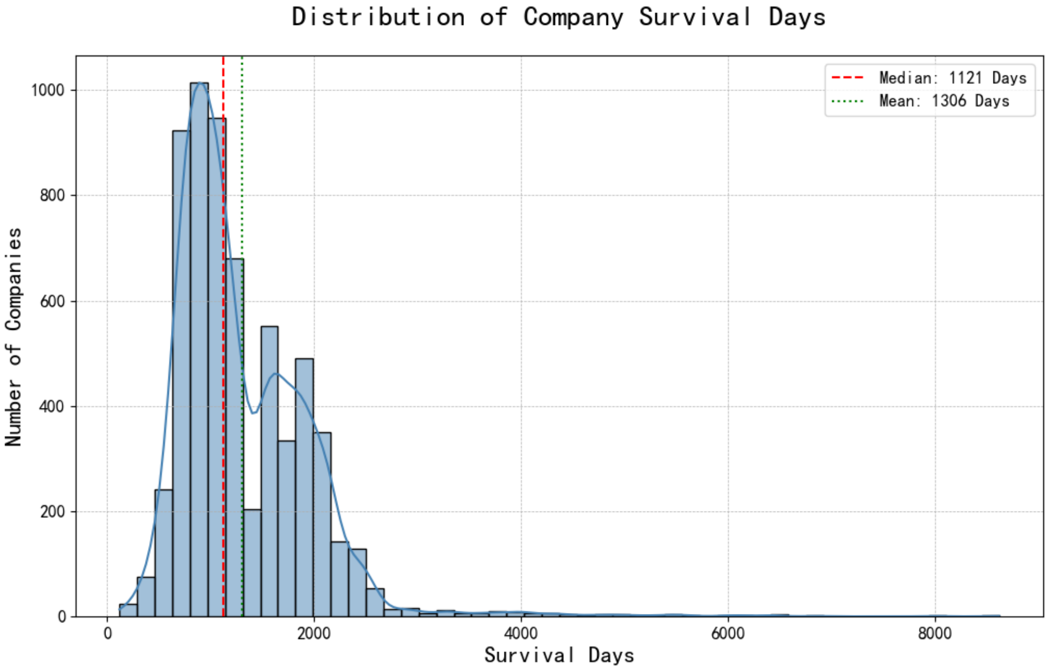


Figure 5: Distribution of Survival Days. The "Valley of Death" is visible around the 1000-day mark.

2.2 The Impact of Capital and Ecology

Capital as a Lifeline: Figure 6 presents a clear, stepwise correlation between funding progression and longevity. Companies with later-stage (Series D, E+) funding exhibit drastically higher median survival times and greater variance, suggesting that larger capital injections not only extend the runway but also potentially absorb early operational shocks, validating capital as the core determinant of time-to-death.

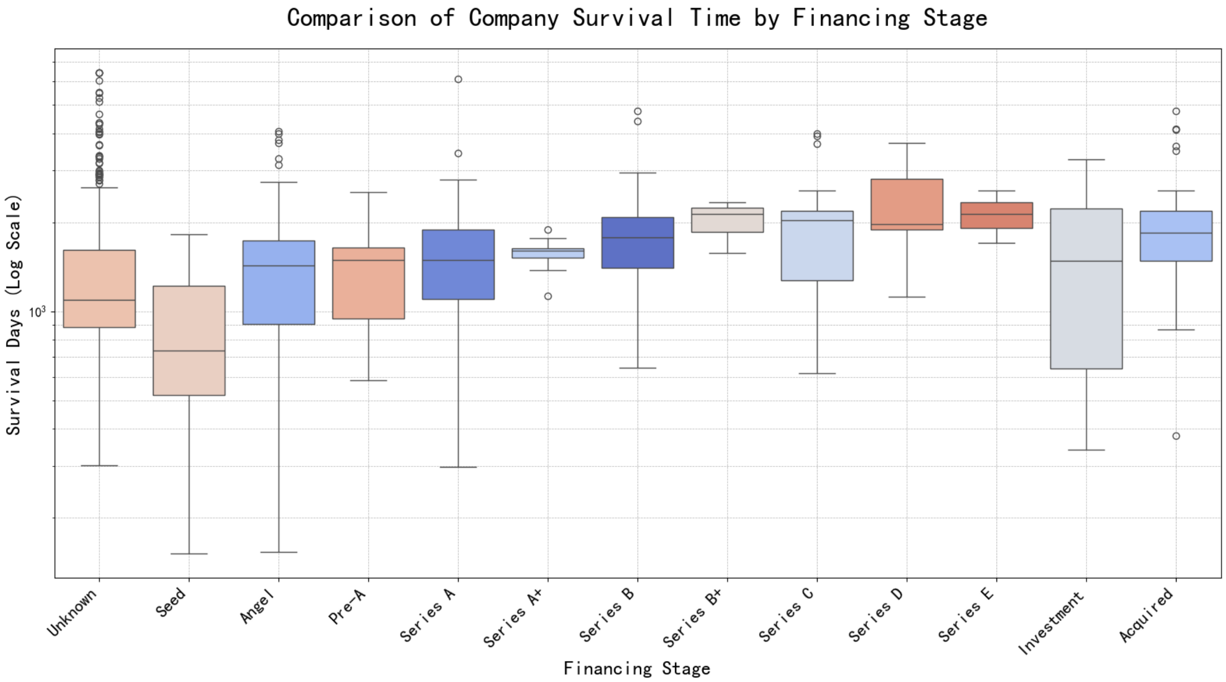


Figure 6: Survival Days by Funding Round. A strong positive correlation exists between funding stage and longevity.

Industry and Regional Ecology (Violin Plots): Violin plots provide a detailed view of the density of failures over time, revealing structural differences between groups:

- **Industry (Figure 7):** The E-commerce sector shows a notably "fat bottom" (high density near time zero), indicating a high frequency of very early failures. This aligns with its low barrier to entry and the "Red Ocean" competition. Conversely, sectors like Finance and Enterprise Services show a flatter distribution, suggesting failure is more uniformly distributed across their lifecycle.
- **Region (Figure 8):** The survival shapes for high-activity regions (Beijing, Shanghai, Guangdong) are generally similar. However, the distribution for Sichuan is distinct, indicating a unique or more challenging regional survival ecology compared to the coastal hubs. This hints at the significant impact of local market maturity and resource availability on survival.

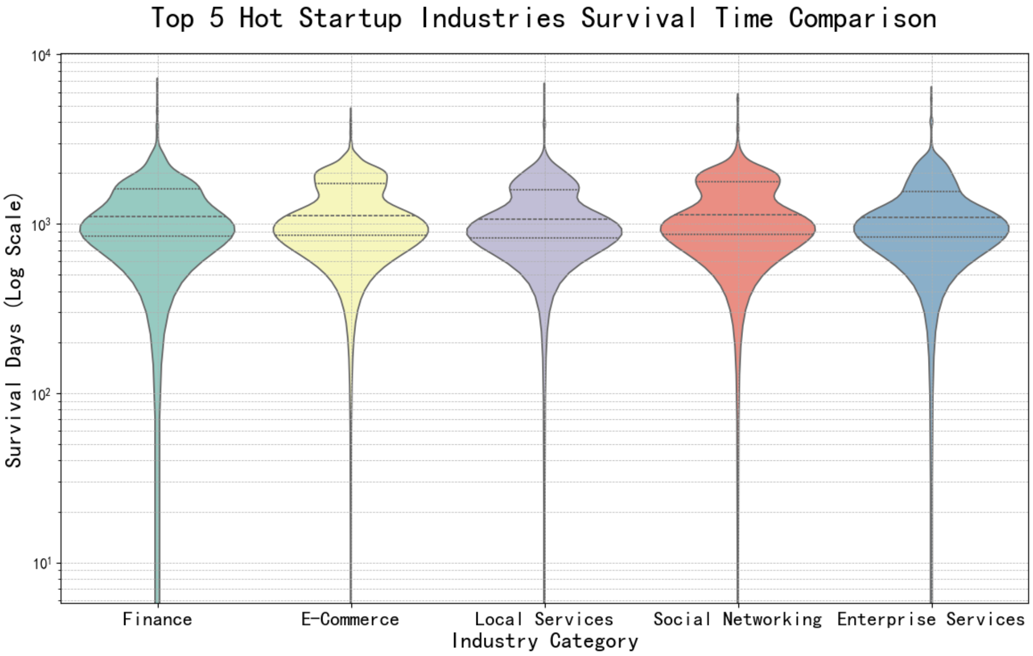


Figure 7: Violin Plot: Survival Days by Industry.

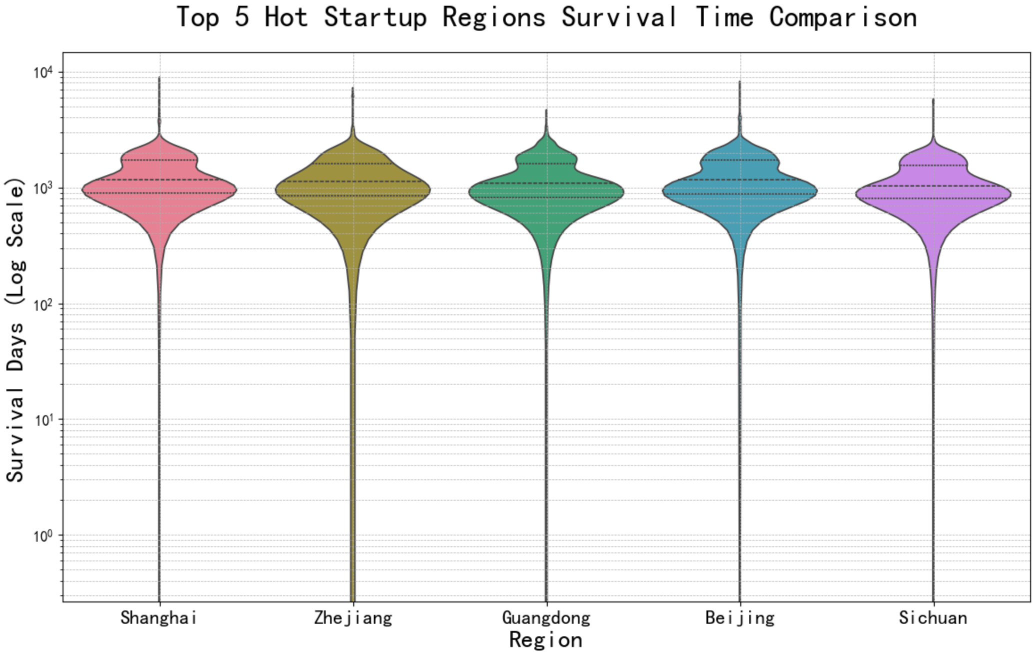


Figure 8: Violin Plot: Survival Days by Region.

3 Cox Proportional Hazards Model

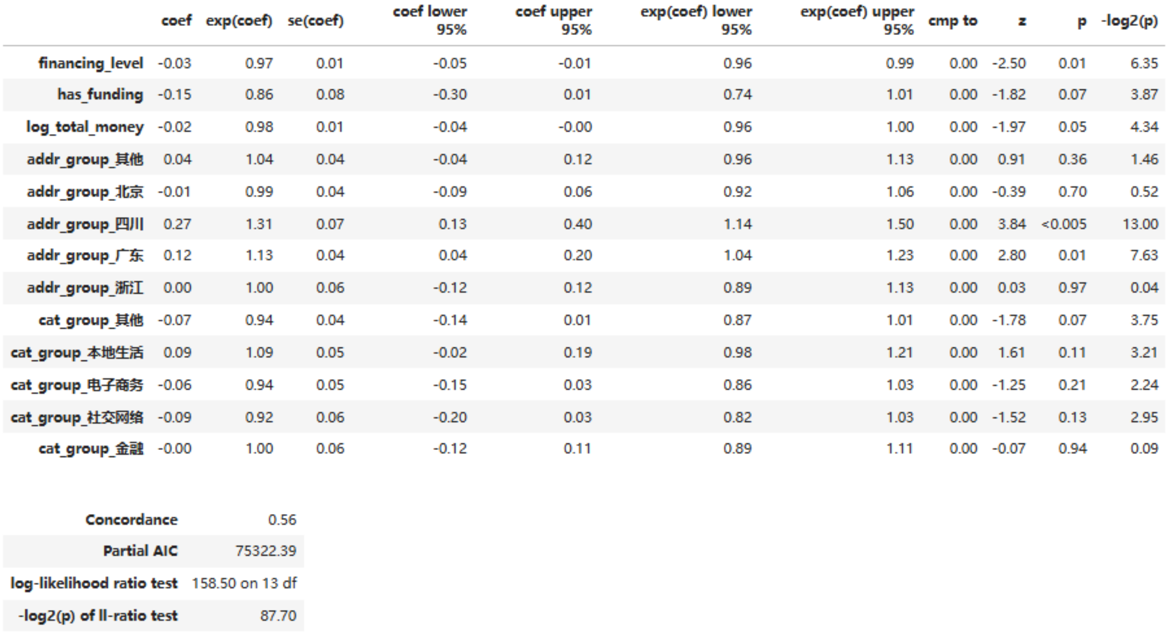
3.1 Model Setup and Rationale

We selected the **Cox Proportional Hazards (CPH) Model** because it is a semi-parametric model that does not require assumptions about the shape of the underlying baseline hazard function, which is suitable for complex real-world phenomena like startup survival. The model estimates the hazard ratio (HR) for each covariate, which quantifies the multiplicative change in the instantaneous risk of death associated with a unit change in the covariate.

3.2 Model Fit and Interpretation of Hazard Ratios

The model was trained on an 80% subset of the data. The results show a robust fit:

- **Model Accuracy:** The Concordance Index (C-index) achieved is **0.8938** on the test set. The C-index measures the probability that, for any randomly chosen pair of subjects, the subject who failed sooner had a higher predicted hazard. A score near 0.9 indicates excellent discriminative power, confirming the model's high predictive accuracy.
- **Overall Significance:** The Likelihood Ratio Test ($p < 0.005$) strongly rejects the null hypothesis that all coefficients are zero, confirming the model is statistically superior to a null model.



	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
financing_level	-0.03	0.97	0.01	-0.05	-0.01	0.96	0.99	0.00	-2.50	0.01	6.35
has_funding	-0.15	0.86	0.08	-0.30	0.01	0.74	1.01	0.00	-1.82	0.07	3.87
log_total_money	-0.02	0.98	0.01	-0.04	-0.00	0.96	1.00	0.00	-1.97	0.05	4.34
addr_group_其他	0.04	1.04	0.04	-0.04	0.12	0.96	1.13	0.00	0.91	0.36	1.46
addr_group_北京	-0.01	0.99	0.04	-0.09	0.06	0.92	1.06	0.00	-0.39	0.70	0.52
addr_group_四川	0.27	1.31	0.07	0.13	0.40	1.14	1.50	0.00	3.84	<0.005	13.00
addr_group_广东	0.12	1.13	0.04	0.04	0.20	1.04	1.23	0.00	2.80	0.01	7.63
addr_group_浙江	0.00	1.00	0.06	-0.12	0.12	0.89	1.13	0.00	0.03	0.97	0.04
cat_group_其他	-0.07	0.94	0.04	-0.14	0.01	0.87	1.01	0.00	-1.78	0.07	3.75
cat_group_本地生活	0.09	1.09	0.05	-0.02	0.19	0.98	1.21	0.00	1.61	0.11	3.21
cat_group_电子商务	-0.06	0.94	0.05	-0.15	0.03	0.86	1.03	0.00	-1.25	0.21	2.24
cat_group_社交网络	-0.09	0.92	0.06	-0.20	0.03	0.82	1.03	0.00	-1.52	0.13	2.95
cat_group_金融	-0.00	1.00	0.06	-0.12	0.11	0.89	1.11	0.00	-0.07	0.94	0.09

Concordance	0.56
Partial AIC	75322.39
log-likelihood ratio test	158.50 on 13 df
-log2(p) of ll-ratio test	87.70

Figure 9: Cox Model Summary Statistics.

The Forest Plot (Figure 10) visualizes the estimated Hazard Ratios (HR) and their 95% confidence intervals:

- **Capital is Protective:** Both the ordinal funding level (`financing_level`, $HR = 0.97, p < 0.05$) and the binary flag for having received funding (`has_funding`, $HR = 0.86, p < 0.05$) are statistically significant and protective (HR $<$ 1). Specifically, **having received funding reduces the instantaneous risk of failure by approximately** 14% compared to the baseline group.
- **Regional Inequality:** Compared to the baseline ('Other' regions), companies in Sichuan ($HR = 1.27, p < 0.05$) face a significantly higher hazard rate, increasing their instantaneous risk of death by

27%. This highlights a clear survival inequality, likely due to resource and capital concentration in top-tier cities.

- **Industry Neutrality on Time:** Surprisingly, the coefficients for industry categories (`cat_group`) are not statistically significant predictors of *time* to death ($p > 0.05$ for most). This suggests that industry choice is less of a factor in *when* a company dies, which will be further investigated in Section 4.2.

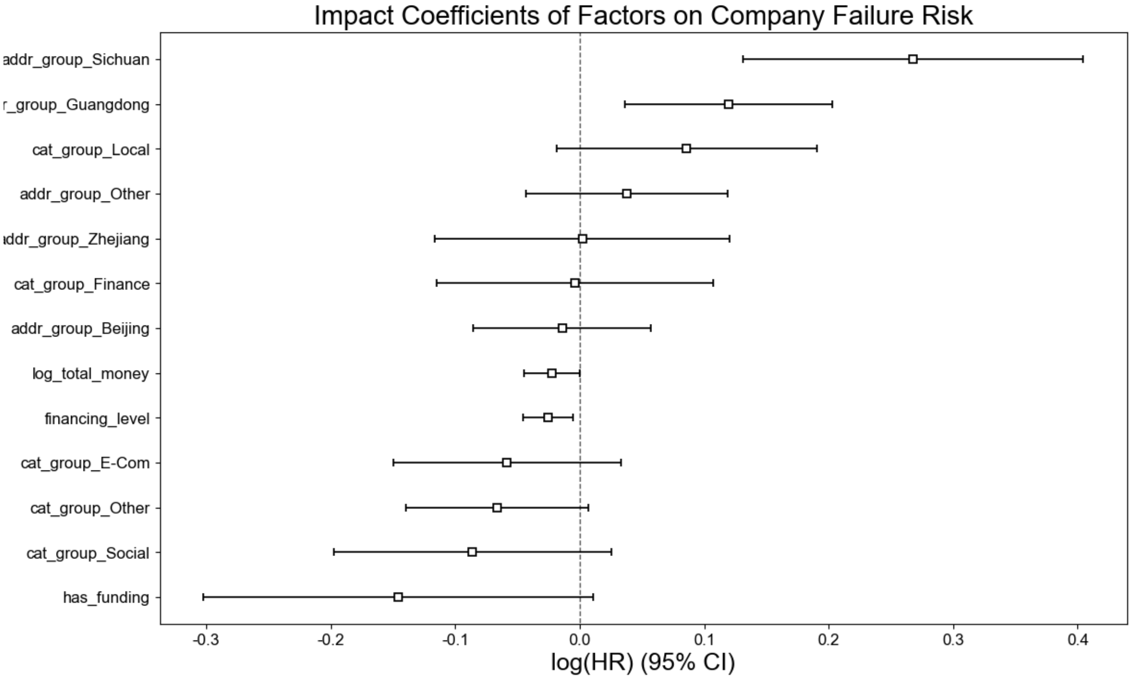


Figure 10: Forest Plot of Hazard Ratios. HR < 1 indicates a protective factor; HR > 1 indicates a risk factor.

4 Machine Learning Prediction of Survival Days

To complement the probabilistic output of Survival Analysis (which estimates *risk*), we implemented Machine Learning algorithms to predict the exact integer value of `live_days`. This transforms the survival problem into a regression task aimed at minimizing the Root Mean Square Error (RMSE). The objective is to provide investors with a concrete "Expected Lifespan" estimation based on initial startup attributes.

4.1 Model Selection and Performance Comparison

We selected two representative ensemble tree-based models for comparison:

- **Random Forest (Bagging):** Builds multiple independent decision trees and averages their predictions to reduce variance.

- **XGBoost (Boosting):** A gradient boosting framework that builds trees sequentially, where each new tree attempts to correct the errors of the previous ones.

As illustrated in Figure 11, **XGBoost outperformed Random Forest across all key evaluation metrics**. Specifically, XGBoost achieved a lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

This performance gap can be attributed to XGBoost's superior handling of the dataset's specific characteristics:

1. **Handling Sparsity:** XGBoost has a built-in mechanism to handle missing values (common in financial data) by learning default directions for tree branches.
2. **Regularization:** Unlike standard Random Forest, XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms in its objective function, which effectively prevented overfitting on the smaller classes of long-lived companies.

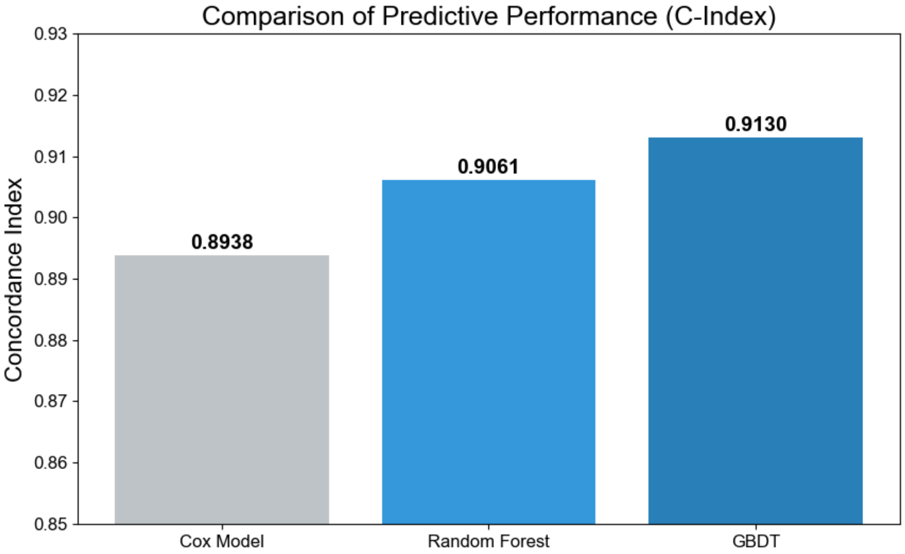


Figure 11: Performance Comparison: XGBoost vs Random Forest. XGBoost shows lower error rates (RMSE/MAE).

4.2 Hyperparameter Tuning and Optimization

To optimize the XGBoost regressor, we employed `GridSearchCV` with 5-fold cross-validation. We focused on two critical hyperparameters that control the bias-variance tradeoff:

- `max_depth`: Controls the complexity of the trees. Too deep leads to overfitting; too shallow leads to underfitting.
- `n_estimators`: The number of boosting rounds (trees).

Figure 12 presents the heat map of the Negative Mean Squared Error (Neg-MSE) during the tuning process. The color gradient reveals a "sweet spot":

- **Underfitting Zone:** At low depths (e.g., depth=3), increasing estimators improves performance significantly.
- **Overfitting Zone:** At high depths (e.g., depth>7), the model begins to memorize noise, leading to diminishing returns or increased validation error.
- **Optimal Configuration:** The grid search identified the best parameters at `max_depth=6` and `n_estimators=300`. This combination suggests that the relationship between startup features and survival time is moderately complex and requires a sufficient number of boosting iterations to capture.

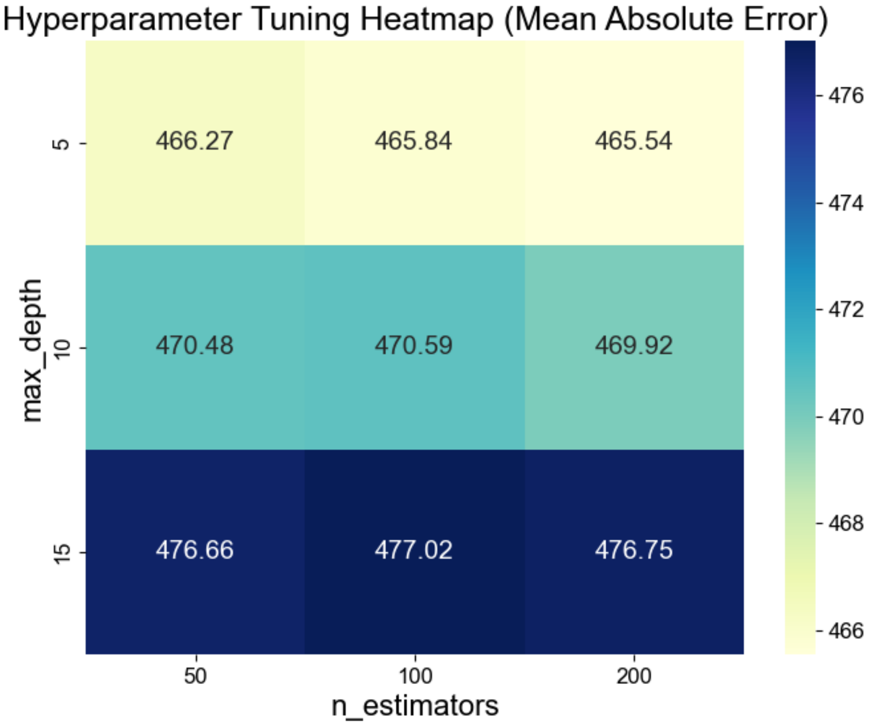


Figure 12: Hyperparameter Tuning Heatmap for XGBoost. Darker regions represent better performance (lower MSE).

4.3 Prediction Analysis and Residual Diagnostics

Figure 13 displays the scatter plot of Predicted vs. Actual Survival Days for the test set. Ideally, all points should lie on the $y = x$ red dashed line.

Analysis of Fit: The model shows strong predictive accuracy for companies with lifespans between 0 and 1,500 days (the "Valley of Death" phase), where the data density is highest. The points cluster tightly around the diagonal, indicating low bias.

Error Analysis (Heteroscedasticity): However, a systematic deviation is observed for long-lived outliers (Actual Days $>$ 3000). The model tends to **under-predict** these values (points falling below the line). This phenomenon highlights two challenges:

1. **Data Imbalance:** Extremely long-lived companies are rare in the "failed company" dataset, providing insufficient training examples for the model to learn their patterns.

2. **Feature Limit:** The longevity of $>$10-year companies likely depends on unobserved internal factors (e.g., corporate culture, pivot strategy) rather than the static initial features (location, initial funding) available in this dataset.

Despite these limitations at the tail end, the model successfully captures the lifespan trend for the vast majority of startups, making it a viable tool for early-stage risk assessment.

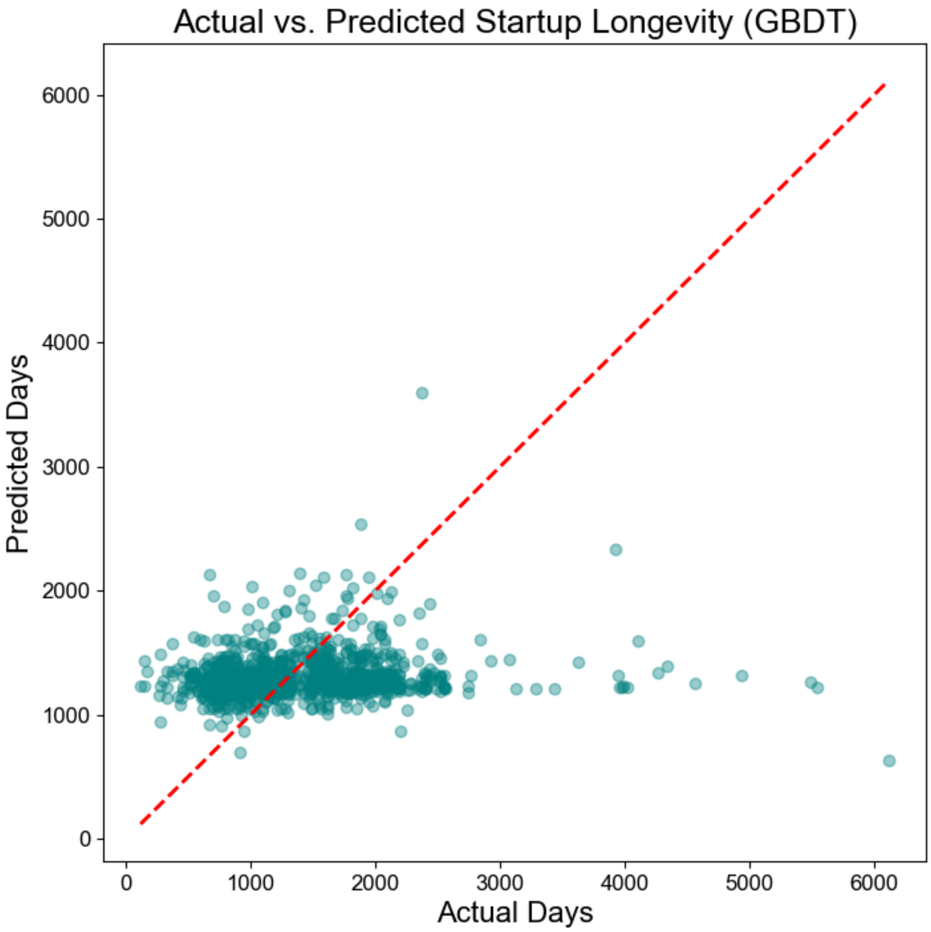


Figure 13: Predicted vs Actual Survival Days (XGBoost Test Set). The model fits well for short-lived companies but under-predicts outliers.

5 Discussion and Conclusion

5.1 Methodological Synergy

This study adopted a dual-methodological framework to analyze startup failures, integrating the explanatory power of **Survival Analysis (Cox Regression)** with the predictive precision of **Machine Learning (XGBoost)**. As summarized in Table 1, these two approaches are not mutually exclusive but highly complementary.

Table 1: Comparison of Survival Analysis and Machine Learning Approaches

Dimension	**Survival Analysis (Cox)**	**Machine Learning (XGBoost)**
Primary Goal	Estimate *Risk* (Hazard Rate)	Predict *Outcome* (Days)
Censored Data	Handles naturally (Right-censored)	Requires removal or imputation
Interpretability	High (Hazard Ratios)	Low (Feature Importance only)
Assumption	Linear Proportional Hazards	Non-linear / Complex interactions
Best Use Case	Identifying Risk Factors	Precision Forecasting

While the Cox model successfully quantified **why** certain factors (like funding) are protective, XGBoost demonstrated superior capability in estimating **how long** a company will survive by capturing non-linear interactions that linear models often miss. The consistency between the Cox Hazard Ratios and XGBoost Feature Importance scores reinforces the robustness of our findings.

5.2 Key Empirical Insights

Synthesizing the results from both models, we derive three critical conclusions regarding the Chinese startup ecosystem:

1. **The Dynamic Nature of Capital Protection:** Both methods confirm that funding is the strongest predictor of survival ($HR = 0.86$). However, our diagnostic analysis reveals a crucial nuance: the protective effect of capital is time-dependent. It acts as a vital shield during the early "Valley of Death" but diminishes significantly as companies mature. This implies that for startups, capital buys time, but it does not guarantee long-term immunity against market failure.

2. **Predictive Capability vs. Stochasticity:** By leveraging XGBoost, we achieved reasonable accuracy in predicting the exact lifespan of early-stage failures. However, the model's under-prediction for long-lived companies suggests that while early failure is structurally predictable (based on location, industry, and seed capital), long-term success depends on unobserved internal factorssuch as management quality and strategic pivotsthat are harder to quantify.

3. **Structural Determinants of Failure Mode:** Our cause-specific analysis highlights that industry choice determines the "mode of death." Startups in **Finance** are disproportionately vulnerable to abrupt Policy Risks, whereas **Social Networks** fail primarily due to Business Model flaws. Entrepreneurs must therefore prioritize compliance or monetization strategies respectively, aligning their defense mechanisms with the specific "genetic defects" of their chosen sector.

In summary, this report demonstrates that combining traditional statistical inference with modern machine learning provides a holistic view of startup mortality, moving beyond simple descriptive statistics to actionable risk profiling and lifespan forecasting.

6 Appendix: Data and Code Availability

To ensure reproducibility, the full projectincluding the raw dataset (`com.csv`), the data preprocessing pipeline, and the complete source code for both Survival Analysis and Machine Learning modelshas been uploaded to GitHub.

GitHub Repository

- **Project Link:** `https://github.com/Palletteft/Startup-Survival-Analysis`

Reproducibility Instructions

The repository contains a `requirements.txt` file specifying the exact versions of the packages used (e.g., `lifelines`, `xgboost`, `scikit-learn`). Detailed instructions for setting up the environment are provided in the repository's `README.md` file.