Bankruptcy prediction and analysis

Based on *Dot-com Bubble* and *GFC* data from the Taiwan Economic Journal for the years 1999 – 2009

Problem Statement

The goal of this project is to understand the financial components that could explain the bankruptcy of a company, especially during bubbles or financial crashes.

Specifically, the goal is to understand in detail how companies' financial data really correlated with how well it is actually doing; get a deeper understanding on how these two crashes affected the economy; be able to predict if a company will go bankrupt based on its publicly available financial data; and understand how such machine learning model works through AI explainability to know what to look out for.

The idea was to, using these insights, check which were the most bankruptcy-prone companies we all know and love from today's public markets. However, as will be discussed in the Dataset Overview section, that won't be explicitly computed because of unknown dataset scales and lack-of-extrapolation of data from 20 years ago to the current market.

This project is approached from the perspective of a personal, for-fun project; but it can have numerous business applications, from trading and investment firms to minority investors, since the information unveiled here might have impactful economic potential. The insights from this project should not be taken as financial advice.

Dataset Overview

The dataset used, namely the Taiwanese Bankruptcy Prediction Dataset [1] collects data from the Taiwan Stock Exchange between the years 1999 and 2009. Company bankruptcy was defined based on the business regulations of such exchange. It contains 95 features, going from simple market metrics such as "Value Per Share" to complex calculated features such as "Long-term fund suitability ratio". Each entry contains one company (anonymized), if it was labeled as bankrupt, and its other 94 features (sadly, no dates).

One problem with this dataset —and the reason why it cannot be extrapolated to the current market— is that some features are normalized in the range [0-1], but we do not know the normalization metric used (which are the minimum and maximum values used by a MinMaxScaler), so our insights are sometimes bounded by "high/low in relation to other datapoints" rather than knowing the exact metric or threshold. That makes it impossible to input current market values to the model, but the insight we can get from it is still very valuable. It also contains a fairly high amount of outliers, however the final decision was to keep them in the analysis, since even giants can go bankrupt.

Some Exploratory Data Analysis is available in the Figures and Charts subsection in the Appendix.

Questions and Objectives

The main questions that will be answered during this project are:

- How did the market look in 1999 to 2009?
- Which features indicate high company health and stability?
- Which features are red flags for bankruptcy?

Methodology

To kickstart the project, the first thing to do is to do some basic exploratory data analysis, though since there are too many variables, at first is kind of hard to visualize properly. After that, some models will be trained (RandomForest, SVM, LogisticRegression, XGBoost), to see which one behaves best.

For the best trained model, apply *SHAP* explainability to understand how the model is behaving internally. After understanding the most important features, iterate again to do a more refined EDA for better results and visualizations. Also, the same models are trained with a reduced set of important features to see the effect of not having them.

Finally, a Streamlit application will be developed to visualize everything dynamically, with live demonstrations.

Another DEMO with today's markets was planned, but after seeing the problems with the dataset and the inability to dodge them, it is not included in the final project delivery.

Background

The Dot-com Bubble (late 1990s – 2000) was a speculative stock market frenzy that peaked in March 2000, centered on technology and internet-based companies ("dot-coms"). The direct causes were a speculative fever where investors ignored traditional financial metrics, pouring capital into companies with little to no revenue, driven by the "growth over profits" mindset and the perceived potential of the internet. Low interest rates in the late 1990s provided abundant capital. This speculative environment meant that companies' financial data was effectively ignored; valuations (market capitalization) bore no relation to fundamental metrics like net income or cash flow. The crash exposed the folly of companies with high "burn rates" and non-existent profits, leading to the NASDAQ Composite index falling nearly 78% by 2002. ([2] [3])

The 2008 Global Financial Crisis (GFC) was a worldwide economic collapse triggered by the implosion of the US housing market. The direct causes were excessive speculation on property values fueled by predatory subprime lending (high-risk mortgages). These mortgages were packaged into opaque, complex financial instruments, primarily Mortgage-Backed Securities (MBS), which were held by financial institutions globally. The failure of these assets, coupled with the banks' extreme overleveraging (high debt-to-equity ratios), led to a systemic loss of confidence and the bankruptcy of Lehman Brothers. For financial institutions, the crisis was fundamentally about financial data integrity: the toxic

assets were wildly overvalued on their balance sheets, and when the housing market fell, the resulting write-downs eroded their capital, revealing widespread solvency and liquidity crises. ([2] [4])

Both crises had severe but distinct implications worldwide (US, Europe, Taiwan, etc.), and were deeply correlated with financial assets and metrics that were overlooked. In the next section, we will discuss which are the most important things to look out for in a company.

Conclusions

Based on the most accurate trained model (Logistic Regression), the three most important features that predict if a company will go bankrupt or not are:

- **Net income to total assets**: ratio between how much the company earns and at how much is it valued. It is a measure of efficiency. A higher percentage indicates better asset utilization company health.
- **Persistent EPS in the last four seasons**: earnings per share (not driven by one-off events) maintained over the last year.
- **Debt ratio**: proportion of a company's assets that are financed by debt. It indicates the company's degree of leverage and financial risk. A higher ratio means the company relies more on debt to fund its assets and may face greater difficulty during economic downturns, whereas a lower ratio suggests greater financial stability.

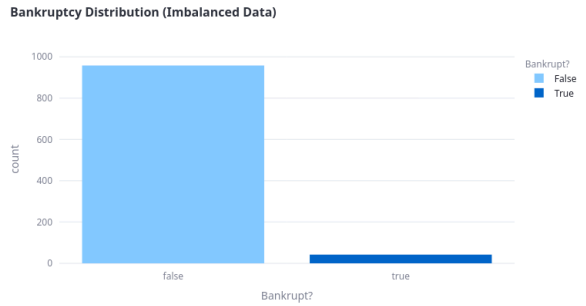
On the one hand, if a company has low net income to total assets' ratio, low or negative persistent EPS in the last four seasons, and high debt ratio; it is very much in trouble, and quite likely to go bankrupt during a bubble burst.

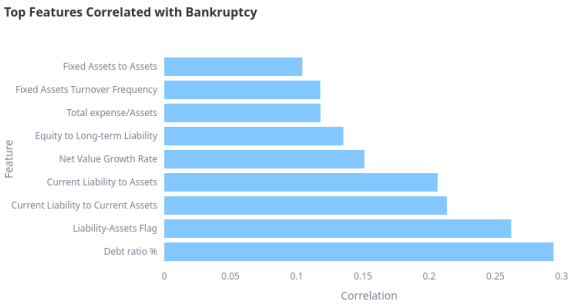
On the other hand, if a company has high net income to total assets, high negative persistent EPS in the last four seasons, and low debt ratio; it is a sign of a very healthy company, that probably won’t go bankrupt in the short term.

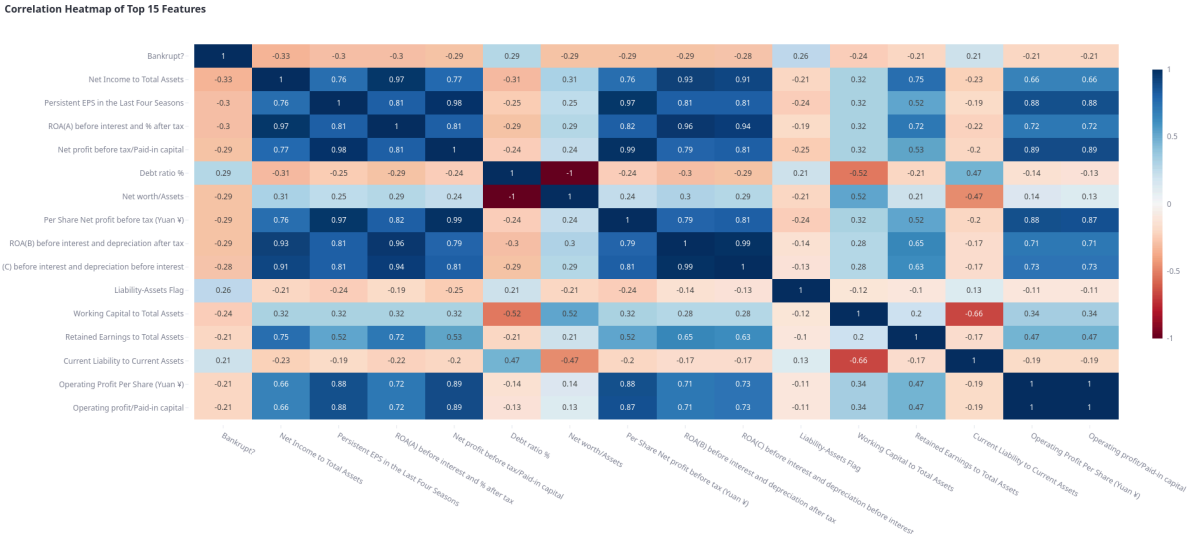
These are only three of 94 feature predictors, and most of them have some kind of weight in the prediction, however solely based on these three we can already see how those can impact today’s high-growth technology and AI companies, especially when balance sheets are obscured by complex financial instruments and commitments.

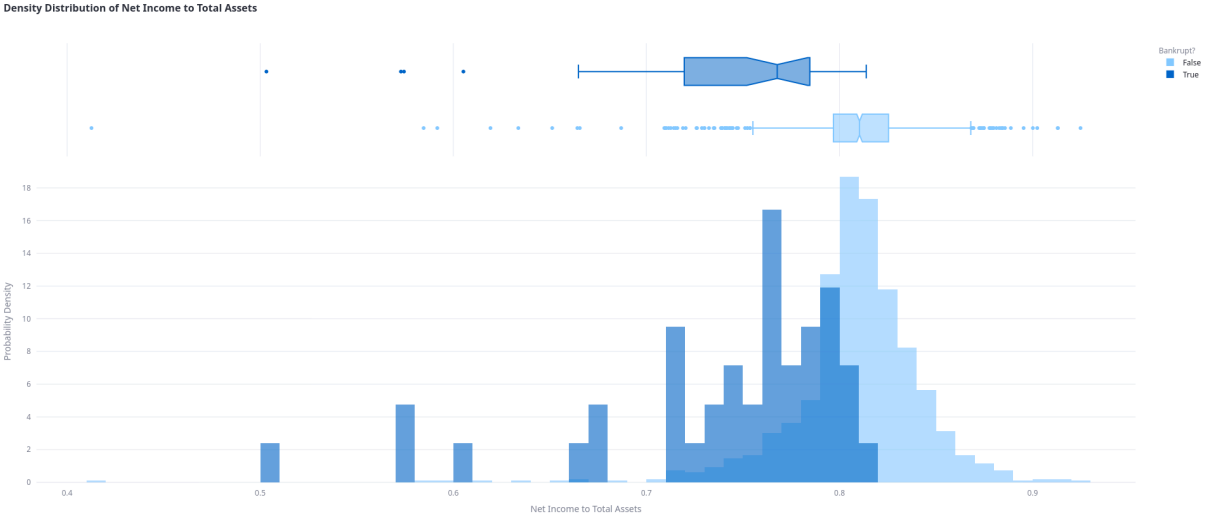
Appendix

Figures and charts

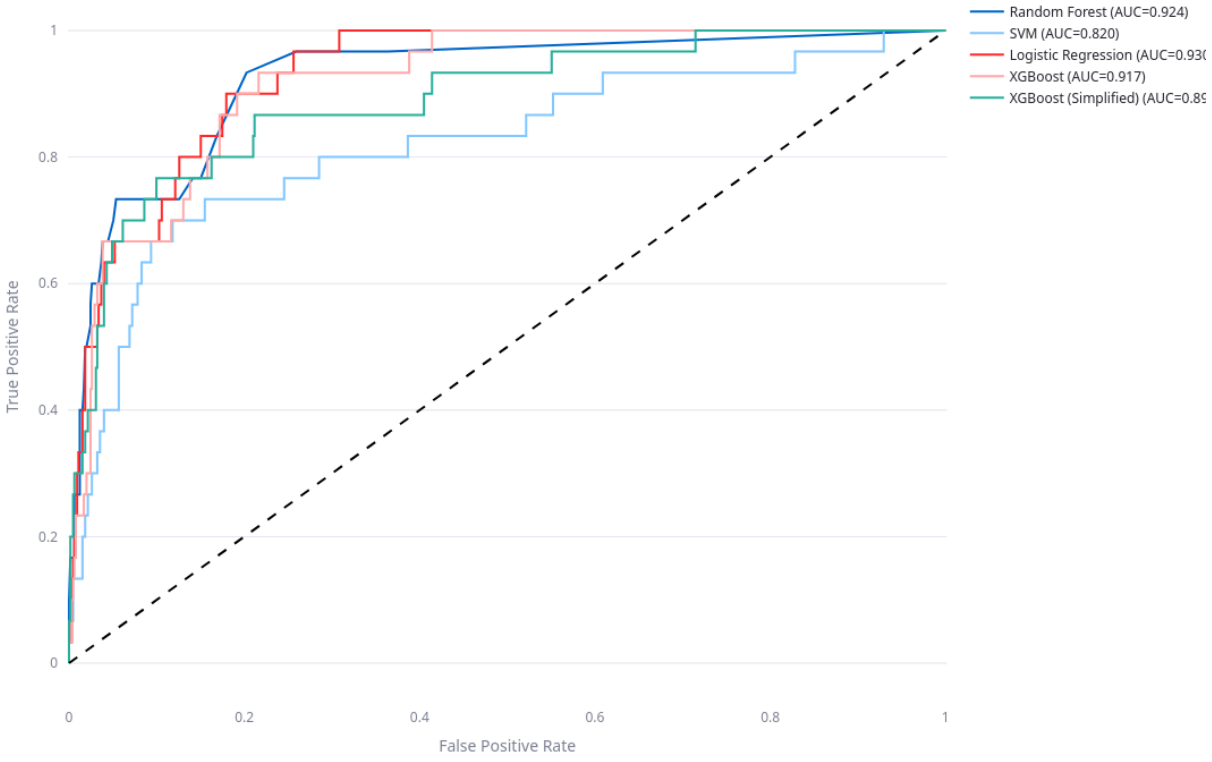




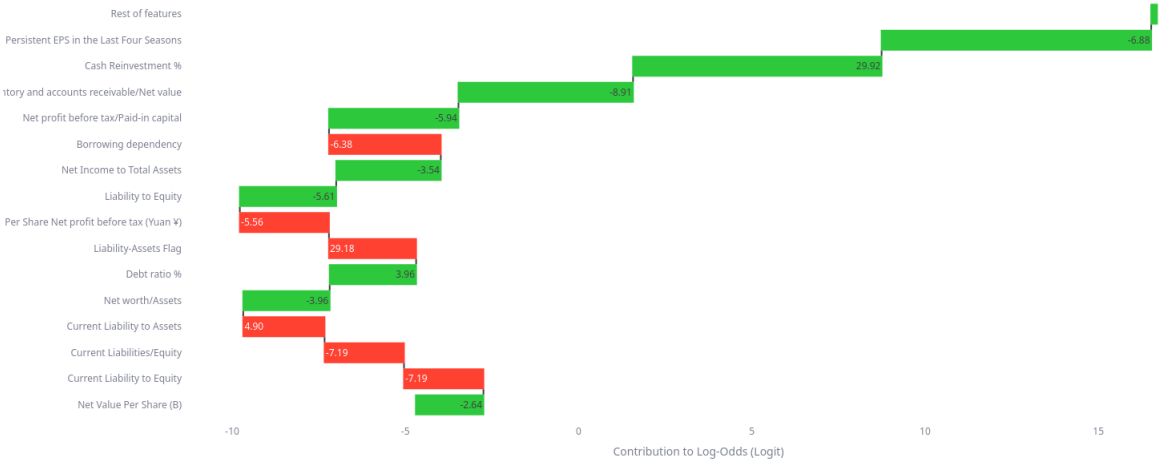




ROC Curve Comparison



Waterfall Chart (Logit Scale) | Final Prob: 1.0000



References

[1] "Taiwanese Bankruptcy Prediction," UCI Machine Learning Repository, 2020. [Online]. Available: https://doi.org/10.24432/C5004D.

[2] W. H. Lau, “2000-2009 US stock market: From dotcom bubble to subprime crisis,” INPressInternational, https://www.inpressinternational.com/post/2000-2009-us-stock-market-from-dotcom-bubble-to-subprime-crisis (accessed Dec. 6, 2025).

[3] “Dot-com bubble,” Wikipedia, https://en.wikipedia.org/wiki/Dot-com_bubble (accessed Dec. 6, 2025).

[4] “2008 financial crisis,” Wikipedia, https://en.wikipedia.org/wiki/2008_financial_crisis (accessed Dec. 6, 2025).