Predicting US Company Bankruptcy Using Accounting Data

Victoria Lu ‘24, Data Science Major Capstone

References:
Dataset: https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset

T., & Curry, B. (2023, October 17). Forbes.
https://www.forbes.com/advisor/investing/fed-funds-rate-history/

Research Questions and Background

Research questions

- How accurately can company bankruptcy be predicted?
- What factors best contribute to predicting company bankruptcy?

Why is this question important?

Answering this question is important to...

- banks when it comes to extending credit to companies
- investors who want to invest in a company’s stock or bonds
- companies themselves as a way detect distress early
- the broader market, as rising bankruptcy occurrences often signal financial stress and tight economic conditions

Extensions to STAT 228 project

- STAT 228 project-Taiwan company bankruptcy prediction
- Experimentation with more data balancing techniques
- Additional time/economic cycle dimension
- Exploration of interaction model and fitting separate models for different cycles

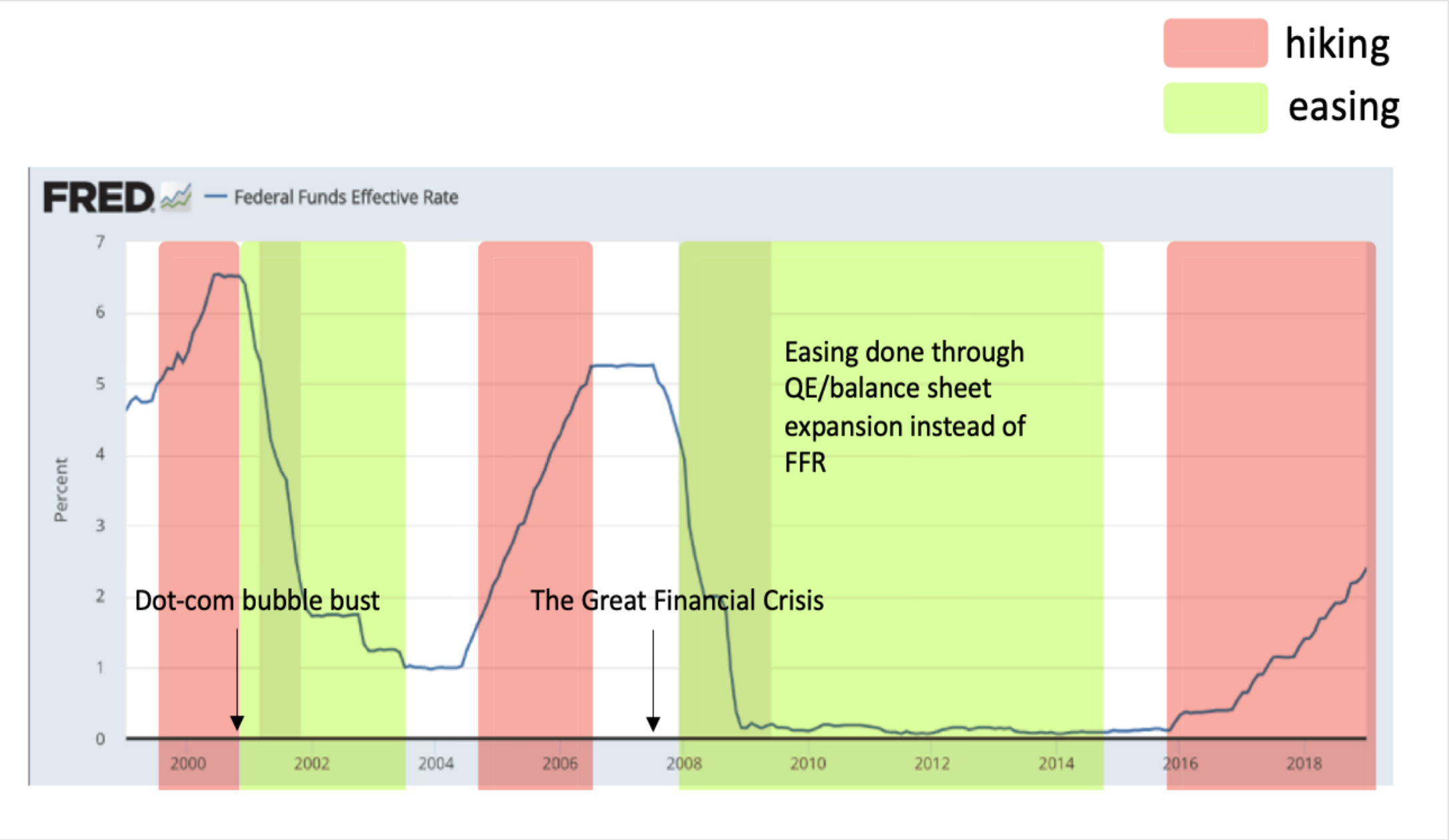
Data Cleaning

Overview of Dataset

- US Company Bankruptcy (Kaggle)
- The dataset contains accounting data from 8,262 distinct companies recorded during the period spanning from 1999 to 2018, which sums to a total of 78,682 observations of firm-year combinations.
- Response of interest: status_label (=1 if alive, = 0 if bankrupt)
- Predictors: 20 numerical accounting variables in millions (e.g., gross profit)
- No missing or duplicated values

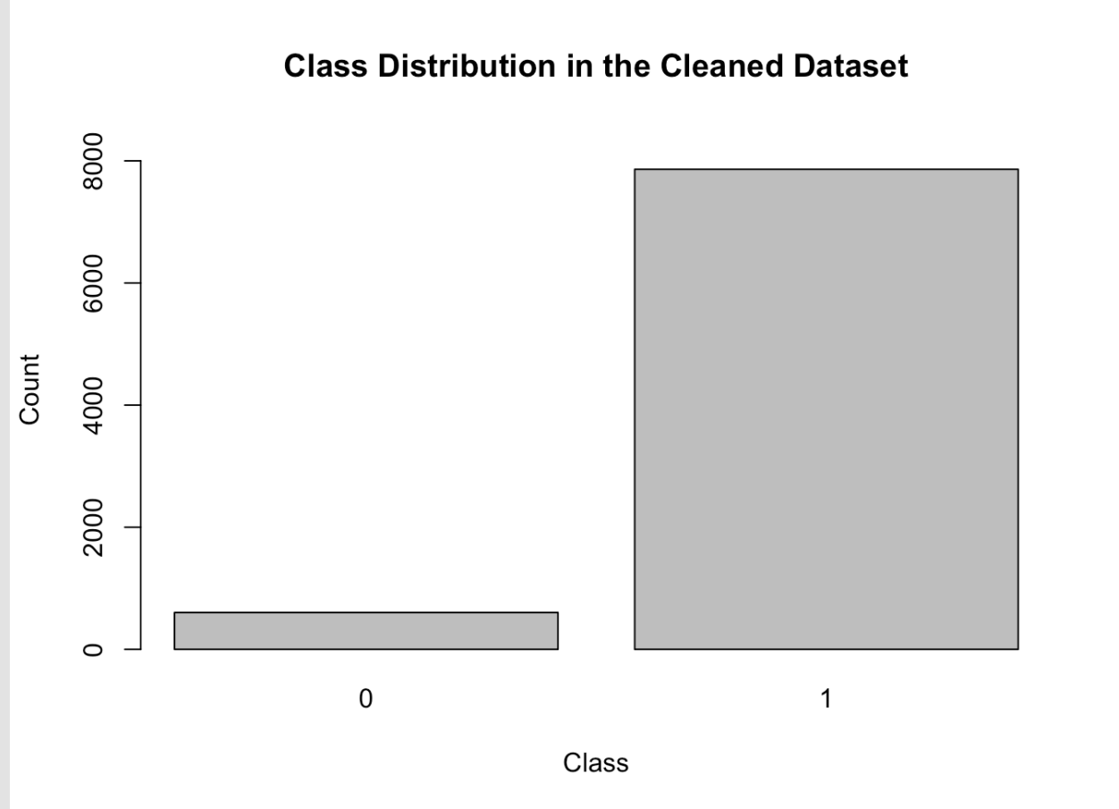
Modifications to the Data

- Cross-sectionalized the data by randomly sampling one observation for each distinct company (78,682 >> 8,262 observations)
- Added a new categorical variable called cycle_type (=1 if year belongs to a hiking cycle, = 0 if year belongs to an easing cycle; ambiguous years were dropped; see Figure below).
- Calculated 20 financial (2 liquidity, 4 profitability, 4 leverage, 6 efficiency, and 4 other) ratios from the accounting variables to account for company size.



Handling Data Imbalance

The cleaned dataset is highly imbalanced with 603 minority classes (=0, bankrupt) and 7862 majority classes (=1, alive). For the testing set, I sampled 10% of the minority class and 10% of the majority class. I made the rest of the observations the training set.



	Minority class (=0)	Majority class (=1)
Training	543	7076
Testing	60	786

To balance the training set, I experimented with undersampling, oversampling, synthetic minority oversampling (SMOTE) with kNN where k=5, and ensemble technique. I evaluated the performance of a bagging classifier. SMOTE yields the best performance overall, so I proceeded with the SMOTE training set.

	Sensitivity	Specificity	Misclassification rate
Unbalanced (543 minority 7076 majority)	0.9975	0.01667	0.0721
Undersampled (543 minority 543 majority)	0.5776	0.6500	0.4173
Oversampled (7076 minority 7076 majority)	0.9669	0.1333	0.09220
SMOTE with 5 nearest neighbors (7076 minority 7076 majority)	0.9262	0.3000	0.11820
Ensemble methods (3 balanced datasets each consisting of 543 minority and 543 majority labels)	0.5369	0.3000	0.45863

Models-SVC & SVM

I tuned the parameters cost and gamma and fitted an SVC, SVM with a polynomial kernel, and SVM with a radial kernel. Their performance metrics are summarized below. SVM with a radial kernel has the best performance overall,

	Sensitivity	Specificity	Misclassification rate
SVC (cost = 1)	0.4389	0.6500	0.5461
SVM with polynomial kernel (cost = 0.1)	0.0140	0.9833	0.9173
SVM with radial kernel (cost = 450, gamma = 40)	0.8104	0.3500	0.2222

Models-Logistic Regression

VIF Screening & PCA

- I started by checking for multicollinearity. VIF screening identified that there are 5 variables with multicollinearity issues.
- I proceeded to handle this via two approaches: 1) Dropping the 5 variables and 2) Using PCA.

Feature Selection

- For each approach, I used AIC and BIC stepwise selection, resulting in four first-order models (logreg.aic, logreg.bic, logreg.aic.pca, and logreg.bic.pca).

Model Evaluation & Diagnostics

- The models have a similar performance. I chose the BIC model without PCA to be the best first-order model given that it is the most parsimonious and interpretable.
- Based on the diagnostic plots for the model, I identified 6 outliers. Removing them resulted in some minor changes in the models.

Interaction Model

- I further selected for interaction terms based on the BIC model without PCA. The interaction model did not show a significant improvement in performance. There are also no patterns in what kinds of interaction terms were added.

The performance metrics of the all models are summarized below.

	Sensitivity	Specificity	Misclassification rate
AIC model without PCA	0.6921	0.4833	0.3227
BIC model without PCA	0.6959	0.4667	0.3203
AIC model with PCA	0.6858	0.4833	0.3286
BIC model with PCA	0.6858	0.4833	0.3286
BIC model without PCA with interaction effects	0.5611	0.5833	0.4374

Model Interpretation

logit(status_label)=-0.5343 +0.0492*Current_Ratio
-0.0002*Return_on_Assets
+0.0000022*Inventory_Turnover_Ratio
+0.0005209*Accounts_Payable_Turnover_Ratio
+1.432*cycle_type1

- Out of the features selected for, two of them are efficiency ratios, one is a liquidity ratio, and one is a profitability ratio.
- Most of the coefficient signs are as expected (e.g. the positive coefficient on Current_Ratio suggests that a higher Current_Ratio is associated with a higher likelihood of a company being alive).
- However, it’s surprising to find that a higher Return_on_Assets is associated with a lower likelihood of a company being alive.
- It also seems counterintuitive that companies in a hiking cycle are more likely to be alive compared to companies in an easing cycle. A possible explanation for this is that the effects of monetary tightening are lagged and that more bankruptcies is what causes the Federal Reserve to ease.

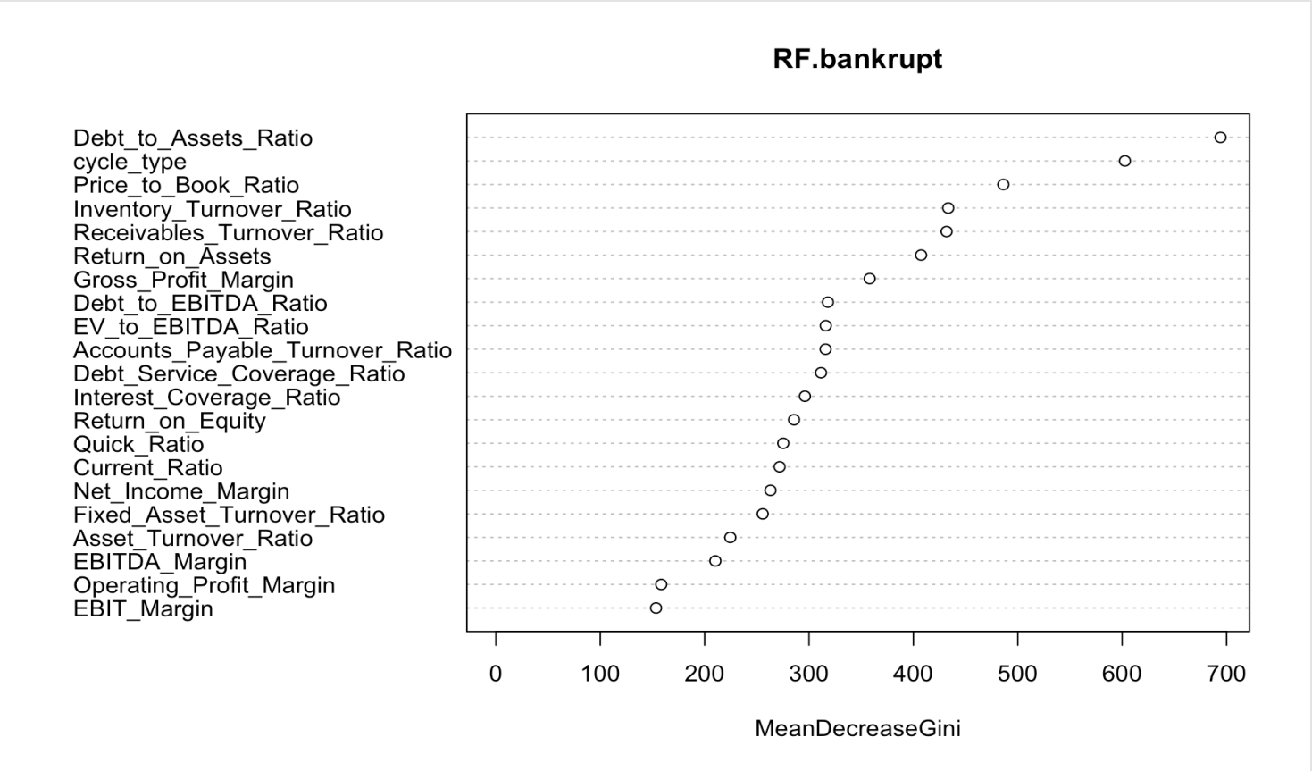
Models-Tree-Based Methods

I fitted a CART, bagging, and random forest model on the training set. Their performance metrics are summarized below. Random forest has the best performance overall.

	Sensitivity	Specificity	Misclassification rate
CART	0.7646	0.4500	0.2577
Bagging	0.9288	0.2333	0.1206
Random forest	0.9173	0.2833	0.1277

Interpretation

- The importance plot below shows that cycle_type, Debt_to_Asset, Price_to_Book, and Inventory_Turnover_Ratio are the most important predictors of bankruptcy.
- Interestingly, Debt_to_Asset and Price_to_Book were not selected for in the BIC model without PCA. This may indicate that there exist some non-linear relationships between bankruptcy and Debt_to_Asset or Price_to_Book that the BIC model was unable to capture.



Fitting separate models for different cycles

- For exploratory and interpretation purposes, I further fitted a separate random forest model for cycle_type=0 and cycle_type=1. Based on the importance plots, the top 5 most important predictors remain the same.
- Nevertheless, it is worth noting that leverage ratios are generally ranked higher during a hiking cycle. This makes sense as higher interest rates impact companies’ ability to borrow and service interest payments.

Best Model & Limitations

Best model-BIC model without PCA

- Comparing the three best models identified in each section, I prefer the BIC model without PCA the most given that it has the highest specificity, which is slightly more important than the other two measures in this context, and that it is the easiest to interpret.

Data Ethics & Limitations

- My project does not have any data ethics concerns as all the financial data are from public companies.
- A lot of time data is still omitted. Conducting time-series analysis would be ideal.
- Despite significant efforts to balance the dataset, the models all have a low specificity. Thus, a better approach may be to start with a more specific and balanced dataset. For example, instead of looking at thousands of companies, many of which may be at a close to zero risk of bankruptcy (e.g., companies like Apple), it may be more helpful to focus on companies that have a rating of below a B- credit rating by the S&P.