# Initial Project Idea

**Financial Modeling** (bankruptcy likelihood indicator using SEC EDGAR filings and other data); Santosh Ananthraman, mentor, see <https://www.linkedin.com/in/santosh-ananthraman-504118/>

**Task:** To build a "company bankruptcy likelihood indicator" using company filings in SEC's EDGAR; select a handful of companies (say, 10) and extract training and test data sets - BBBY *[Bed, Bath, & Beyond]* et al, i.e., a good selection of bankrupt/tending towards bankruptcy/non-bankrupt companies from the recent past (say, 4/2/4 from each category).

**Tools:** Google Colab and Python libraries; results shared in Google Drive.

**Collect the data:** Access the EDGAR database (www.sec.gov/edgar) and download the annual 10-K and quarterly 10-Q filings for the companies you are interested in analyzing. Say 16 consecutive quarters worth of 10-Ks and 10-Qs - use qtrs 1-12 for the training set; use qtrs 13-14 for the test set, and qtrs 15-16 to demonstrate ongoing future monitoring. These filings contain comprehensive financial statements, management discussions, risk factors, and other relevant information. Extract the data. Bulk download from CapitalIQ database (I believe CMU libraries can enable access) may also be possible (which will permit a longer, *walk-forward* test and a larger universe).

**Financial ratio analysis:** Extract financial data from the filings and calculate key financial ratios. Common ratios to consider include liquidity ratios (current ratio, quick ratio), solvency ratios (debt-to-equity ratio, interest coverage ratio), profitability ratios (gross margin, net profit margin), and activity ratios (inventory turnover, accounts receivable turnover). Compare these ratios with industry (or sectors, or peer companies) benchmarks to identify potential red flags.

**Trend analysis:** Examine financial statement trends over time. Look for any significant deterioration or improvement in financial metrics. Declining profitability, increasing debt levels, decreasing liquidity, or a consistent pattern of losses may indicate financial distress.

**Cash flow analysis:** Analyze the cash flow statements provided in the filings. Pay attention to the company's ability to generate positive operating cash flows, manage its investing activities, and finance its operations. Negative cash flows from operations and a reliance on external financing can be warning signs.

**Non-financial indicators:** Consider non-financial factors that may affect a company's financial health. These could include changes in management, industry dynamics, customer trends, competitive positioning, and macroeconomic factors. External sources like industry reports and news articles can provide additional insights. Identify public data sources for this info.

**Perform EDA:** Identify and generate 3-4 insightful visualizations from the above features and dimensionality-reduced combinations thereof.

Develop a representative feature vector composed of a linearly independent selection of the above features (that serve to explain most of the variance in the data) for modeling purposes.

**Perform Risk Assessment via Topic Modeling** (e.g., Latent Dirichlet Allocation): Study the management discussion and analysis (MD&A) section of the filings to understand the company's risks, challenges, and strategies. Look for factors that may impact future performance, such as competition, regulatory changes, technological disruptions, or significant litigation.

**Build and test predictive models:** Compare a couple of statistical and machine learning techniques. Incorporate the financial and non-financial indicators identified in the previous steps as input variables. Train the model using historical data that includes both bankrupt and non-bankrupt companies. Evaluate the model's accuracy against the test set and adjust as needed.

**Regular monitoring** (to test continued operation in a production envt): Continuously update your analysis as new filings become available (i.e., test on qtrs 15-16 from your dataset). Monitor any changes in the company's financial condition or business environment that may impact the bankruptcy risk. Adjust your model and predictions accordingly.

Write a research paper describing the above effort. Use Google Colab/Drive.

# 

# Slack Note on independent variables

It might be useful to compare (1) what's been successfully used in your literature search, and, (2) what's available from your data sources (such as CapitalIQ), with the list below and adapt them to create your initial list of Xs (independent variables) for your training/test datasets.

X1    net profit / total assets

X2    total liabilities / total assets

X3    working capital / total assets

X4    current assets / short-term liabilities

X5    [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] \* 365

X6    retained earnings / total assets

X7    EBIT / total assets

X8    book value of equity / total liabilities

X9    sales / total assets

X10    equity / total assets

X11    (gross profit + extraordinary items + financial expenses) / total assets

X12    gross profit / short-term liabilities

X13    (gross profit + depreciation) / sales

X14    (gross profit + interest) / total assets

X15    (total liabilities \* 365) / (gross profit + depreciation)

X16    (gross profit + depreciation) / total liabilities

X17    total assets / total liabilities

X18    gross profit / total assets

X19    gross profit / sales

X20    (inventory \* 365) / sales

X21    sales (n) / sales (n-1)

X22    profit on operating activities / total assets

X23    net profit / sales

X24    gross profit (in 3 years) / total assets

X25    (equity - share capital) / total assets

X26    (net profit + depreciation) / total liabilities

X27    profit on operating activities / financial expenses

X28    working capital / fixed assets

X29    logarithm of total assets

X30    (total liabilities - cash) / sales

X31    (gross profit + interest) / sales

X32    (current liabilities \* 365) / cost of products sold

X33    operating expenses / short-term liabilities

X34    operating expenses / total liabilities

X35    profit on sales / total assets

X36    total sales / total assets

X37    (current assets - inventories) / long-term liabilities

X38    constant capital / total assets

X39    profit on sales / sales

X40    (current assets - inventory - receivables) / short-term liabilities

X41    total liabilities / ((profit on operating activities + depreciation) \* (12/365))

X42    profit on operating activities / sales

X43    rotation receivables + inventory turnover in days

X44    (receivables \* 365) / sales

X45    net profit / inventory

X46    (current assets - inventory) / short-term liabilities

X47    (inventory \* 365) / cost of products sold

X48    EBITDA (profit on operating activities - depreciation) / total assets

X49    EBITDA (profit on operating activities - depreciation) / sales

X50    current assets / total liabilities

X51    short-term liabilities / total assets

X52    (short-term liabilities \* 365) / cost of products sold)

X53    equity / fixed assets

X54    constant capital / fixed assets

X55    working capital

X56    (sales - cost of products sold) / sales

X57    (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)

X58    total costs /total sales

X59    long-term liabilities / equity

X60    sales / inventory

X61    sales / receivables

X62    (short-term liabilities \*365) / sales

X63    sales / short-term liabilities

X64    sales / fixed assets

# 

# Meeting Notes 11/06/2023

Problem Setup: y=F(X) (a spatiotemporal function fitting exercise)

The goal is to generate a model (F) that indicates the likelihood (y) of the impending bankruptcy of a given company one timescale unit (e.g., quarter/year; year is typical in previous research) into the future, given correlated data (X, comprised of x1, x2, etc.) from today and the immediate past, for that company.

How many companies?

1000 for each class (typical # in previous research)?

What scope?

(Public companies listed on stock exchanges; US NYSE & Nasdaq only; or add European exchanges London SE/Deutsche B/French CAC etc?)

(Across all major sectors)

(Bankruptcy=1 is defined as “moment of delisting from exchange”?)

(Keep 2 balanced “classes” of bankrupt=1 vs non-bankrupt=0 in data sets; typical from previous research)

(build training vs testing vs production testing datasets)

Temporal memory setup for model?

Set up as …X(t-2), X(t-1), X(t) => y(t+1) (timescale for lookahead is 1 qtr or 1 year?)

Or create and employ specific trend capturing x’s based on (d/dt) rates-of-change or higher order d/dts, or “area under the curve” features (integral of x dt)

Data

Need to scale and normalize quantitative x’s so they are comparable between companies – from financial statements & from financial ratios

Can we use any qualitative data? Bags of phrases (a textual NLP approach maybe?) that can help separate these two classes? If yes, then augment into above model.

Correlation matrices and dimensionality reduction

Which x’s to use? Check correlations with y, cross correlations with other x’s, and auto-correlations with themselves in time

Also create linear independence in the x’s – use PCA/other to reduce dimensionality if needed)

GOAL for Sunday 6/18/23:

(1) Build and present your 12 minute talk for next Sunday;

(2) In the background start to construct the TRAIN, TEST & PROD datasets.

What are the candidates for F?

Methods/single/hierarchical/ensemble

What measures do we use to assess the Fs?

ROC(AUC)/Confusion Matrices based on TP,FP,TN,FN

These two things come next, after your datasets are ready.

# Workflow

* Phase 1 | **22/06/2023**
  + Make a List of Companies for the non/bankrupt dataset
  + Create the Dataset from CapitalIQ
  + Try to minimize the number of features [financial ratios] using correlation matrix/ L1 or L2 loss function/ Feature Engineering using Descriptive Statistics/ Feature Cross/ Principal Component Analysis
* Phase 2
  + Develop a Supervised Learning Model using **Gradient Boost** [CatBoost] from Financial Ratios or **Random Forest**
  + Train the model
  + Test the model
* Phase 3 [If the accuracy is below expectation and we have enough time]
  + Repeat Phase 2 with Deep Learning
* Phase 4 [If we have time]
  + Develop an Unsupervised Learning Model [Topic Modeling] using NLP from the filing documents
  + Train the model
  + Test the model
* Phase 5
  + Combine the topic model with the gradient boost model
  + Test the accuracy of the combined model

# Roles

* Create Dataset : Yilin Du & AbdulAziz/ Jin
* Build the model : Abrar & Jin
* Train the model : Abrar & Jin
* Test the model : Abrar & Jin
* Create Visualizations: AbdulAziz
* Write Reports: Yilin & AbdulAziz/Jin
* Create Presentations: Jin

# Tasks | due by Sunday 18 June

* Finish reading PApers
* Start Creating TRAIN, TEST & PROD datasets.
* Finish Presentation

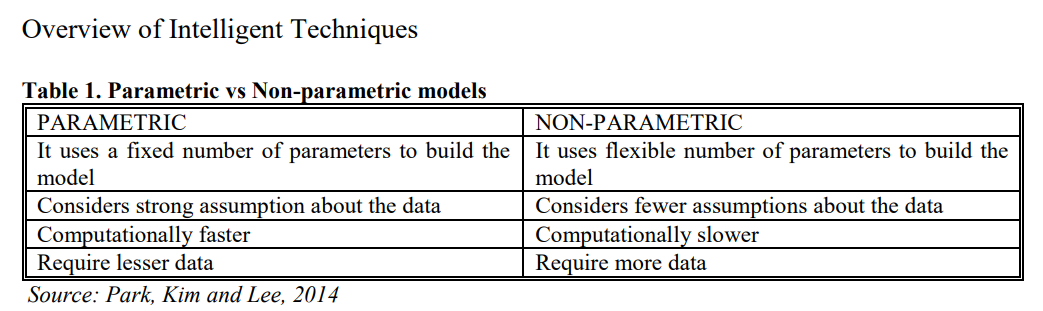
# Papers

Assigned to Everyone

Clement, C. (2020). Machine Learning in Bankruptcy Prediction–a Review. *Journal of Public Administration, Finance and Law*, (17), 178-196.

<https://www.jopafl.com/uploads/issue17/MACHINE_LEARNING_IN_BANKRUPTCY_PREDICTION_A_REVIEW.pdf>

* Parametric vs Non-parametric Models



* Not having equal samples for bankrupt and non-bankrupt companies leads to the well-known oversampling problem where when faced with an imbalance between labels the algorithms tend to predict the oversampled label/class.
* 87,5% of papers used financial ratios as the predicting variables showing a still biased view, in one form or another, towards considering financial ratios as being the only way to go in predicting bankruptcy.
* There has not been found any correlation between the number of features and model accuracy, with models with just two features being just as capable in terms of accuracy as models with 20+ features.

Abrar

Begum, S. (2022). A detailed study for bankruptcy prediction by machine learning technique. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 2* (pp. 201-213). Springer Singapore.

[A Detailed Study for Bankruptcy Prediction by Machine Learning Technique | SpringerLink](https://link.springer.com/chapter/10.1007/978-981-16-6369-7_18)

* Dataset from Taiwan Economic Journal through Kaggle; 7k records with 96 features
* 22 features chosen based on correlation
* The SMOTE Oversampling technique addressed the very low bankrupt-to-non-bankrupt company ratio.
* Random Forest Classifier and ANN tested to perform better than XGBoost & LR

Odom, M. D., & Sharda, R. (1990, June). A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks* (pp. 163-168). IEEE.

[A neural network model for bankruptcy prediction | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/5726669)

* NN’s performance was tested against Discriminant Analysis
* Only 5 financial ratios used
* Classification worked better when trained in a 50/50 BN: NBN company data ratio
* Companies misclassified by NN are also misclassified by Discriminant Analysis.

Brenes, R. F., Johannssen, A., & Chukhrova, N. (2022). An intelligent bankruptcy prediction model using a multilayer perceptron. *Intelligent Systems with Applications*, 200136.

[An intelligent bankruptcy prediction model using a multilayer perceptron - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2667305322000734)

Yilin

1. H. Son, C. Hyun, D. Phan, H.J. Hwang, Data analytic approach for bankruptcy prediction, *Expert Systems with Applications*, Volume 138, 2019, 112816, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.07.033.

the ensemble models: higher prediction accuracy and lower error frequency than statistical models / robust to an imbalance in [data processing](https://www.sciencedirect.com/topics/computer-science/data-analysis) problem

Selection of variables:

Reduce the number of variables for:

1. too many variables cause the incompleteness of data (unavailable data)
2. it is hard to interpret the results if there are many variables.

Not reduce the number of variables for:

Information loss for model performance is unavoidable while removing some variables.

Suggestions for using financial data: reduce skewness

2. Feng Mai, Shaonan Tian, Chihoon Lee, Ling Ma, Deep learning models for bankruptcy prediction using textual disclosures, *European Journal of Operational Research*, Volume 274, Issue 2, 2019, Pages 743-758, ISSN 0377-2217, <https://doi.org/10.1016/j.ejor.2018.10.024>.

Bankruptcy prediction based on textual disclosures:

“The deep learning model using both textual and numeric inputs has improved prediction accuracy over the models using a single type of input.”

“Our results provide direct, large-sample evidence of textual disclosure's information value. The AUC values for data models for the 1-year ahead forecast, depending on the model we use, are between 0.711 and 0.784. When the forecasting horizon is longer (3-year), the predictive value of unstructured text is comparable to audited accounting ratios and market data. Therefore, integrating textual disclosures into risk models could provide great insights.”

3. Shaonan Tian, Yan Yu, Financial ratios and bankruptcy predictions: An international evidence, International Review of Economics & Finance, Volume 51, 2017, Pages 510-526, ISSN 1059-0560, <https://doi.org/10.1016/j.iref.2017.07.025>.

adaptive-LASSO is used to select variables for modeling

For Japan market,

*Retained Earning/Total Asset*

*Total Debt/Total Asset*

*Current Liability/Sales*

The model using those three predictor variables demonstrates statistical significance in bankruptcy prediction over one-year, two-year and three-year-ahead prediction horizons.

4.DOI:[10.5539/ijef.v6n3p1](https://doi.org/10.5539/ijef.v6n3p1)

Zhenghao

<https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0166693&type=printable>

1. They suggest different set of variables for different regions or market. This one can be useful when we have data from different market situations.

2. It mentions that they use different variables for different companies to achieve high accuracy.

3. Ratios they used:

a) financial statements

b) cash flows

c) share performance

d) profitability

e) liquidity

4. ratios of the market and macroeconomics? What is this, worth asking

<https://www.mdpi.com/1911-8074/12/1/30#:~:text=Most%20important%20financial%20ratios%20in%20predicting%20small%20enterprise,share%20of%20net%20financial%20surplus%20in%20total%20liabilities>.

1. Ratios they used:

a) operating profitability of assets

b) current assets turnover

c) capital ratio, coverage of short-term liabilities by equity

d) coverage of fixed assets by equity

e) the share of net financial surplus in total liabilities

2. This is for small enterprises, may be biased to use on large companies

<https://proceedings.neurips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf>

1. This one is a machine learning strategy. It can block out the ratios that have little affect on the result, so that we can have a clean prediction to raise the accuracy.

2. Maybe more useful in deep learning and neuro network, but the idea in this paper is worth considering.

AbdulAziz

<https://studenttheses.uu.nl/bitstream/handle/20.500.12932/26366/Thesis-Frank_Wagenmans-3870154.pdf?sequence=1&isAllowed=y>

* It develops models using Logistic Regression, Neural Network, Random Forest, and Decision Tree.
* The paper demonstrates that models based on payment behavior are capable of predicting bankruptcy in the next 12 months.
* It concludes that Random Forests outperform the other techniques, with Logistic Regression models in conjunction with the Elasticnet regularization technique closely following.
* There's an analysis of the relevance of historical data in bankruptcy prediction and an exploration of how far ahead bankruptcy can be predicted.
* Five ratios: working capital/total assets, retained earnings/total assets, sales/total assets, earnings before interest and taxes/ total assets, market capitalization/ total debt

Fich, E.M., Slezak, S.L. Can corporate governance save distressed firms from bankruptcy? An empirical analysis. *Rev Quant Finan Acc* 30, 225–251 (2008). <https://doi.org/10.1007/s11156-007-0048-5>

* A variety of factors such as board composition, shareholder power, and management structure were investigated for potential correlation with bankruptcy avoidance.
* Robust statistical methods were employed to handle the imbalance in the bankruptcy to non-bankruptcy firm ratio.
* A comparative analysis was conducted among companies with different degrees of corporate governance effectiveness.
* Identified specific governance structures and mechanisms that seem more successful in avoiding bankruptcy.
* Findings were compared with other commonly employed strategies for bankruptcy prediction.
* Ratios used: interest coverage ratio, risk ratio, market-to-book ratio, capital expenditure to sales ratio

<https://www.sciencedirect.com/science/article/abs/pii/S0957417414007660>

* The paper introduces an approach for credit risk evaluation based on linear Support Vector Machines (SVM) classifiers, focusing on applications on larger datasets.
* Utilizes a real-world financial dataset from the SEC EDGAR database for experimental classification performance results.
* The paper addresses the importance of credit risk evaluation and bankruptcy prediction models in the financial domain.
* It notes the rise of machine learning techniques in the credit risk domain, such as artificial neural networks (ANN) and Support Vector Machines (SVM).
* Discusses the advantages of SVM over ANN, such as absence of local minima and relatively simple architecture.
* Despite linear SVM not being extensively explored in this domain, the paper considers it due to its faster performance compared to nonlinear SVM, making it suitable for large-scale SVM classification and regression problems.
* The study aims to explore the potential of linear SVM technique against medium or larger datasets in the credit risk domain, using a "sliding window" approach for training and testing, and optimizing classifier selection using particle swarm optimization.

# 

# Resources

Capital IQ link

<https://guides.library.cmu.edu/az.php?q=capital%20iq&s=100438&p=1>

AutoRegressive Model

<https://en.wikipedia.org/wiki/Autoregressive_model>

Long-short term memory

<https://en.wikipedia.org/wiki/Long_short-term_memory>

AutoCorrelation

<https://en.wikipedia.org/wiki/Autocorrelation>

Confusion Matrix

[Confusion matrix - Wikipedia](https://en.wikipedia.org/wiki/Confusion_matrix)

ARCH time series

[Autoregressive conditional heteroskedasticity - Wikipedia](https://en.wikipedia.org/wiki/Autoregressive_conditional_heteroskedasticity)

Cross-Validation for Time Series

<https://robjhyndman.com/hyndsight/tscv/>

Point-In-Time vs. Lagged Fundamentals

[Microsoft Word - Thought Leadership - PIT vs. Lagged Fundamentals - FINAL (spglobal.com)](https://www.spglobal.com/marketintelligence/en/documents/sp-capitaliq-quantamental-point-in-time-vs-lagged-fundamentals.pdf)

News

[April adds 54 more US corporate bankruptcies; 2023 filings highest since 2010 | S&P Global Market Intelligence (spglobal.com)](https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/april-adds-54-more-us-corporate-bankruptcies-2023-filings-highest-since-2010-75543160)

[List of Retail Company Bankruptcies & Closing Stores | CB Insights Research](https://www.cbinsights.com/research/retail-apocalypse-timeline-infographic/)

[How To Present Research Data? - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4453119/)

Lasso vs Ridge penalty

[Ridge and Lasso Regression: L1 and L2 Regularization | by Saptashwa Bhattacharyya | Towards Data Science](https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b)

CatBoost

<https://catboost.ai/>

Loss Functions

<https://androidkt.com/how-to-add-l1-l2-regularization-in-pytorch-loss-function/>