THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Mining Project: Company Bankruptcy

Amar Adusumilli

Kardelen Cicek

Tarush Gupta

Background and Motivation

In the financial sector, publicly-traded firms are expected to report their annual filings that contain an audit of their financial health, using financial ratios. Additionally, firms maintain corporate governance indicators to satisfy regulatory requirement and oversight.

Prediction using machine learning methods can prevent financial collapses by modeling the parameters that represent a company's bankruptcy probability. It provides a financial insight into a company's health. Using these forecasts, investors and lenders can make appropriate investment decisions in these firms.

Data

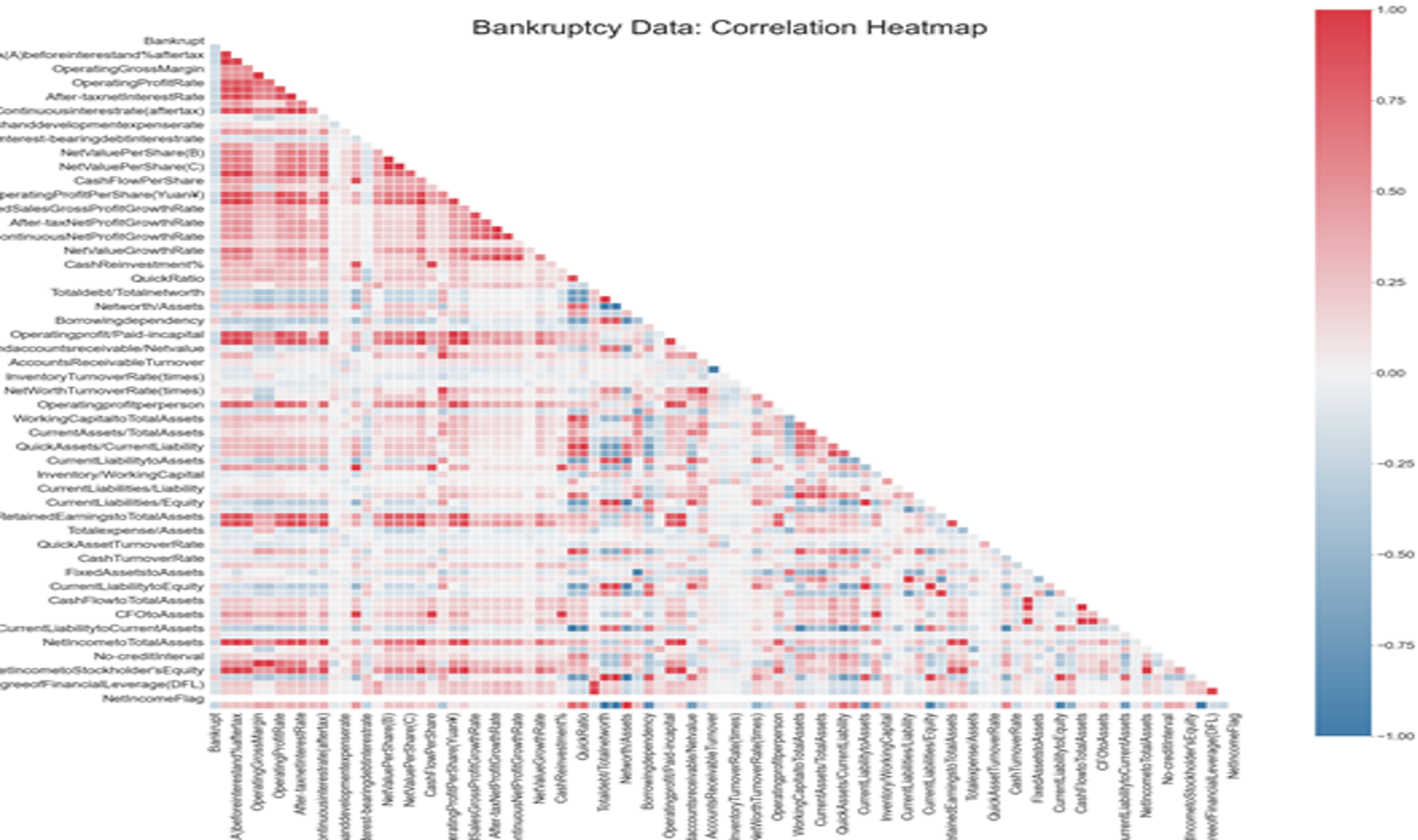
- Kaggle competition: Company Bankruptcy Prediction.
- Data collected from Taiwan Economic Journal, from 1999 to 2009, for firms operating in Taiwan.
- The dataset contains 6,819 rows of observations with 96 columns.
  - One target variable: Bankruptcy indicator, which is of Boolean class.
  - 95 financial health indicators, including financial ratios and corporate governance indicators.
- Well-organized data without missing information.
- Data processed using Python 3.9.

Exploratory Data Analysis

To reduce model complexity and prevent estimation/learning issues, we eliminate highly correlated features.

Spearman's Rank Correlation coefficient matrix asses the relationship between variables.

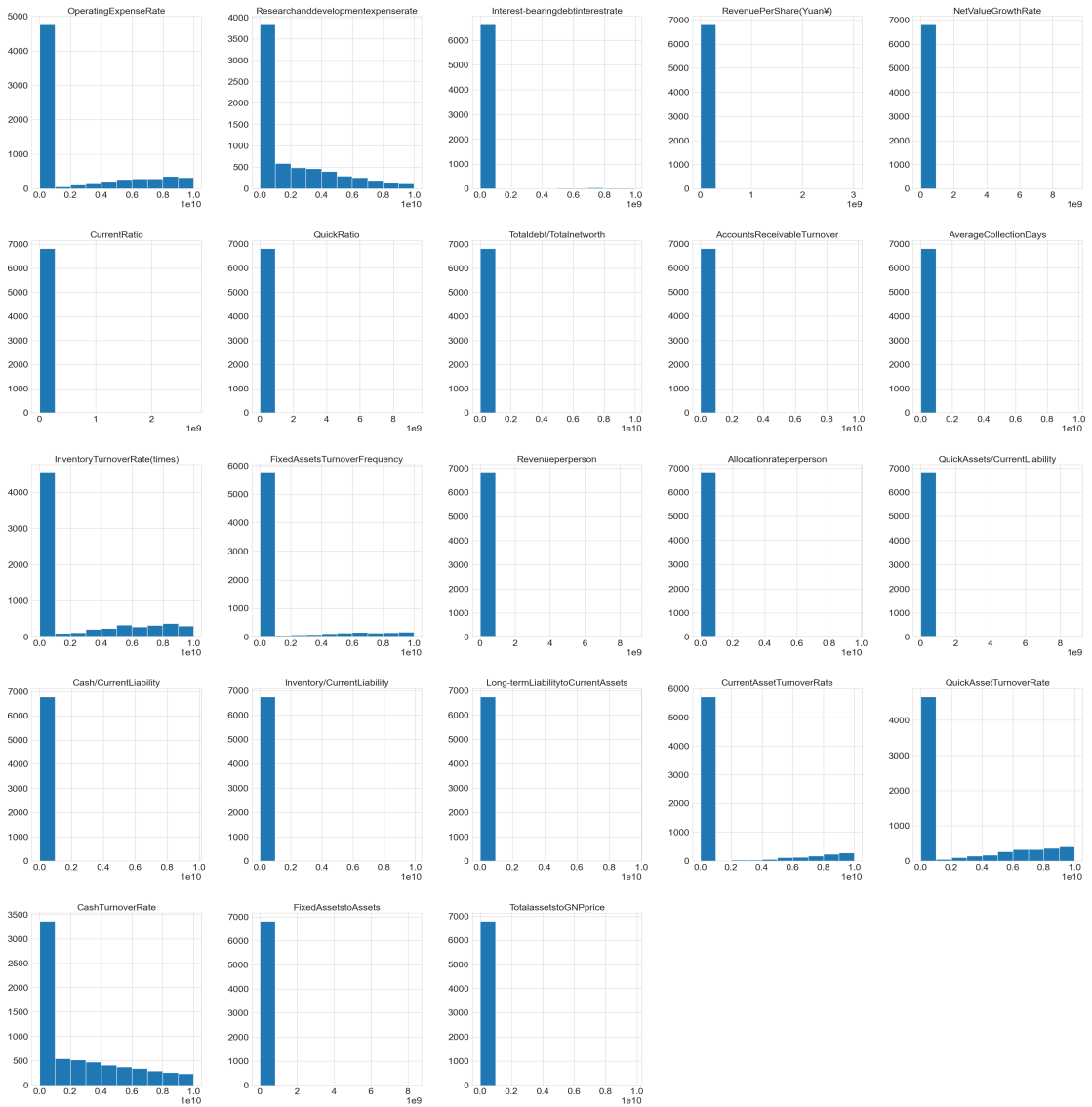
Most features are in the $[0, 1]$ range, however several have very large outliers. For these, we apply Winsorization


Bankruptcy Data: Correlation Heatmap
Bankrupt
OperatingGrossMargin
OperatingProfitRate
NetValuePerShare(B)
NetValuePerShare(C)
CashFlowPerShare
CashReinvestment%
QuickRatio
Totaldebt/Totalnetworth
Networth/Assets
Borrowingdependency
AccountsReceivableTurnover
InventoryTurnoverRate(times)
NetWorthTurnoverRate(times)
WorkingCapitaltoTotalAssets
CurrentAssets/TotalAssets
QuickAssets/CurrentLiability
CurrentLiabilitytoAssets
Inventory/WorkingCapital
CurrentLiabilities/Liability
CurrentLiabilities/Equity
Totalexpense/Assets
QuickAssetTurnoverRate
CashTurnoverRate
FixedAssetstoAssets
CurrentLiabilitytoEquity
CashFlowtoTotalAssets
CFOtoAssets
NetIncometoTotalAssets
No-creditInterval
NetIncomeFlag
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

EDA (Cont.)

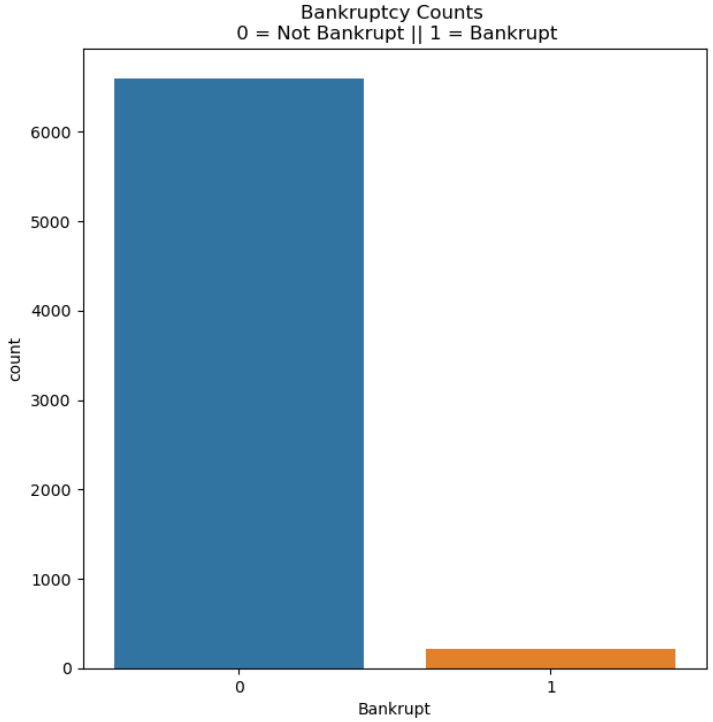
Most features are in the $[0, 1]$ range, however several have very large outliers.

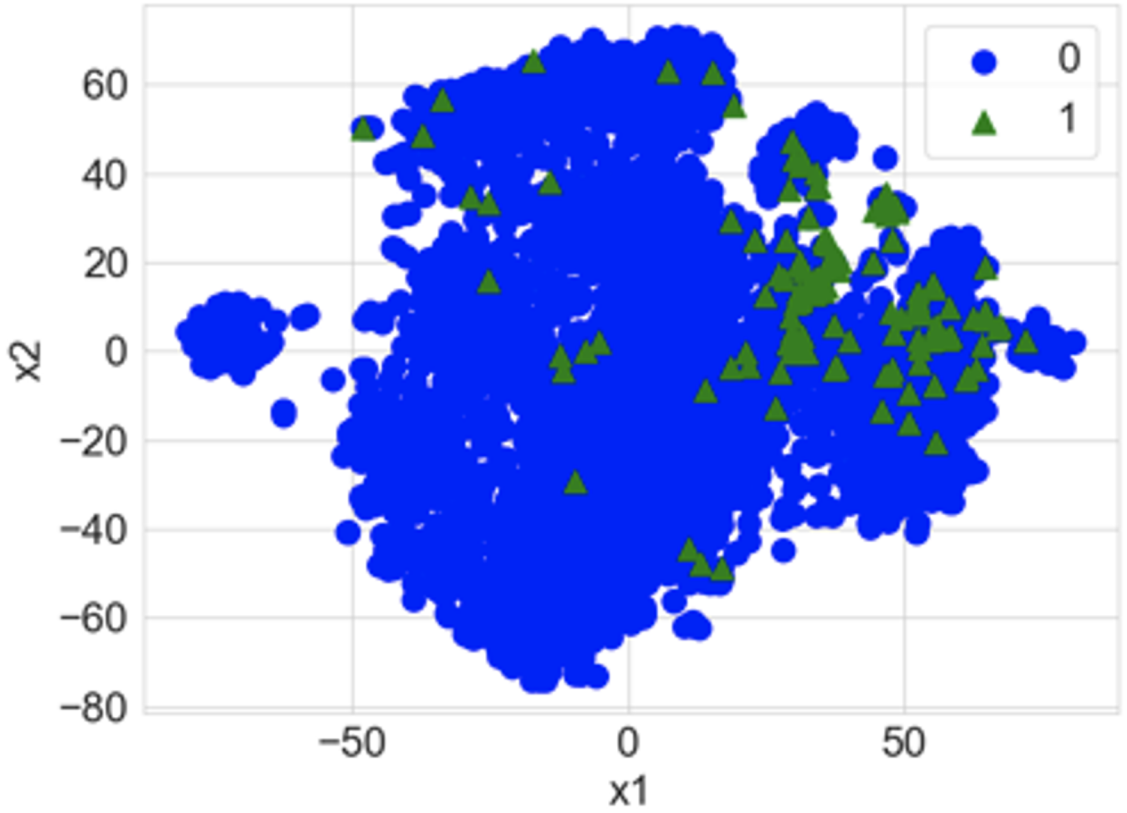
For these, we apply Winsorization

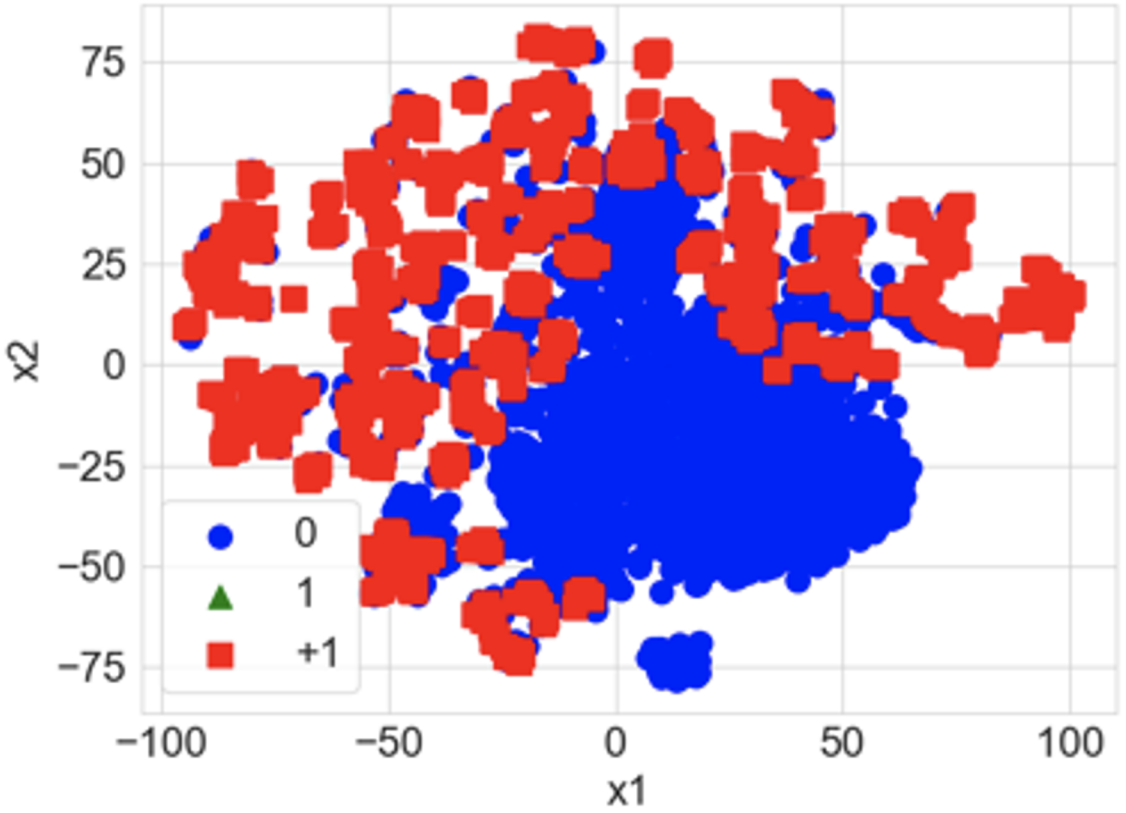


Handling Class Imbalance

- ‘Bankrupt?’: Binary (boolean) target class variable
  - 6,819 rows: 3.2% minority class data
  - Mild imbalance problem
- SMOTE (Chawla et al., 2002)
  - Minority class data randomly selected
  - KNN randomly chosen
  - Generates synthetic minority class data




0
1
60
40
20
0
−20
−40
−60
−80
x2
−50
0
50
x1


0
1
+1
75
50
25
0
−25
−50
−75
x2
−100
−50
0
50
100
x1

Inferential Analysis

- What factors contribute to the likelihood of Bankruptcy?
  - And by how much?

- Two methods:
  - Linear Probability Model
  - Probit Regression
- VIF used to further eliminate highly correlated variables

Linear Probability Model (LPM)

- This is standard OLS applied to a binary dependent variable:

$$y_i = u + \beta_0 + \sum \beta_i x_i$$

- Each coefficient measures the change in the probability that $y = 1$, ceteris paribus

- Not useful for prediction

LPM Results

Variable	Estimate	Standard Error	T-Statistic	P-Value
TotalDebt/TotalNetWorth	0.057	0.006524709	8.725066	2.66E-18
Operatingprofit/Paid-incapital	0.031497086	0.00615543	5.11696	3.11E-07
TotalAssetTurnover	0.020661977	0.005881382	3.513116	0.000443
Totalexpense/Assets	0.01997927	0.005956478	3.354209	0.000796
Cash/CurrentLiability	0.018490886	0.00730577	2.530997	0.011374

Probit Model

- Model is of the form:

$$P(y = 1|x) = G(\beta_0 + \sum \beta_i x_i)$$

- In a Probit model, $G(z) = \Phi(z)$
  - The Standard Normal CDF
  - This transformation ensures an image between $[0, 1]$

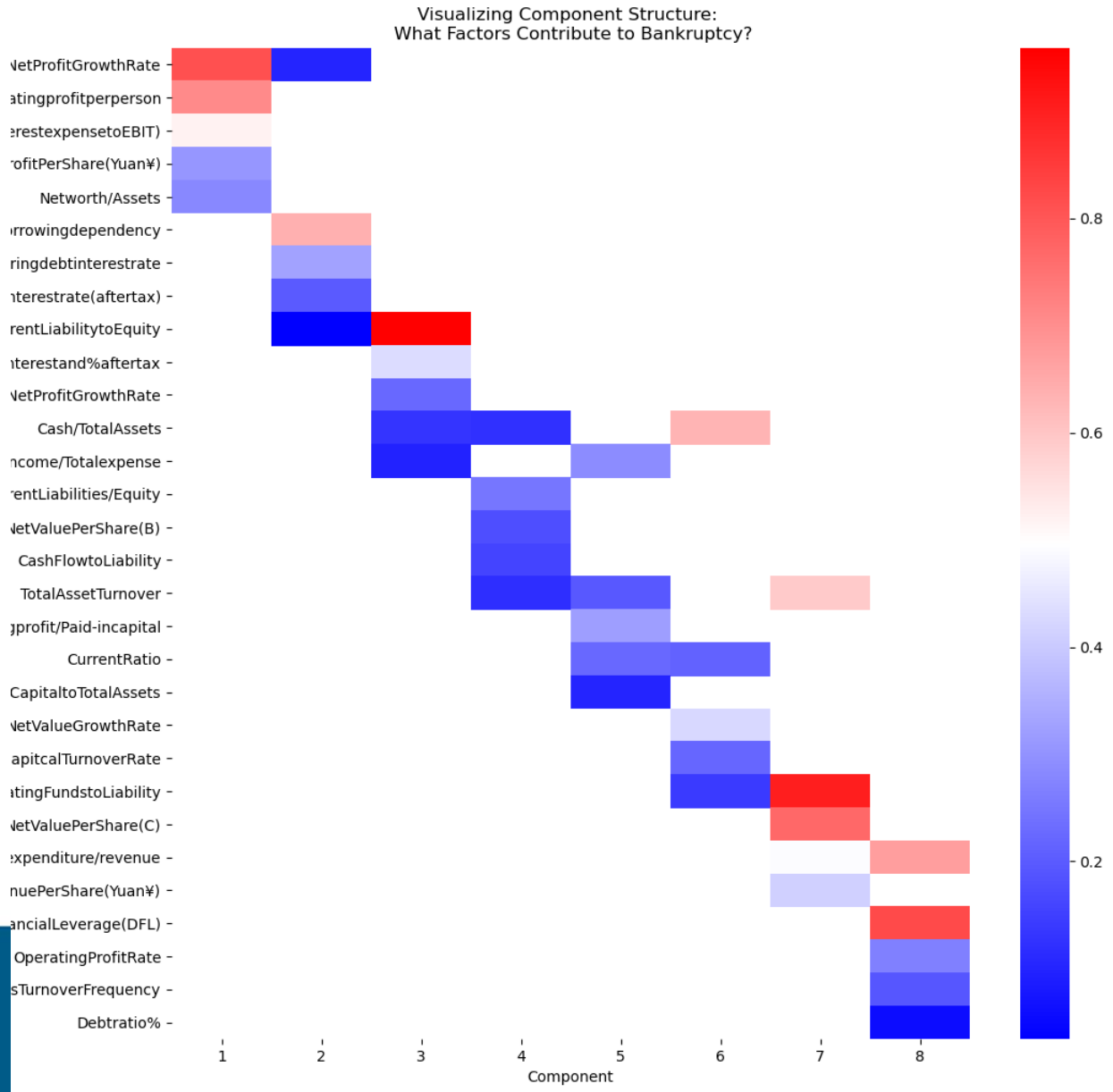
Probit Results

Variable	Parameters	Z-Statistic	P-Value
Operatingprofit/Paid-incapital	0.467751671	3.192098992	0.001412429
Totaldebt/Totalnetworth	0.374387096	7.561067932	4.00E-14
Revenueperperson	0.185454735	3.111594243	0.001860801
GrossProfittoSales	0.097805344	2.107706823	0.035056357

PCA

8 Components explain 97% of variation in the data

Each component is mostly comprised of a set of related variables



Best Model Prediction

Four different models are estimated to predict company bankruptcy data.

1. Decision Tree Classifier: each leaf node assigned as a class label. Root and other internal node contain attribute test conditions to separate records of different characteristics.
2. MLP Classifier: relies on underlying neural network to perform task of classification.
3. Histogram Gradient Boosting Classifier: it is fast and effective for big data sets.
4. Logistic Regression: a method for binary class prediction. It is supervised learning classification algorithm.

Model Prediction Results

Decision Tree Classifier: 0.75, MLP Classifier: 0.74, Histogram Gradient Boosting Classifier:0.74, Logistic regression: 0.61

Best Model to predict company bankruptcy is Decision Tree Classifier with 75 % accuracy.

Bibliography

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589-609.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

Lee, Tsun-Siou, and Yin-Hua Yeh. "Corporate governance and financial distress: Evidence from Taiwan." (2004): 378-388.

Liang, Deron, Chia-Chi Lu, Chih-Fong Tsai, and Guan-An Shih. "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study." *European Journal of Operational Research* 252, no. 2 (2016): 561-572.

Lin, Wei-Yang, Ya-Han Hu, and Chih-Fong Tsai. "Machine learning in financial crisis prediction: a survey." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, no. 4 (2011): 421-436.

Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

Shailer, Gregory EP. *An introduction to corporate governance in Australia*. Pearson Education Australia, 2004.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC