CS 559 Final Project Report

Company Bankruptcy Prediction

Elina Rezaeian - 20019320

Hargun Sawhney - 20015108

Ishaan Chawathe - 20031207

Ottikunta Sahith - 20033091

I pledge my honor that I have abided by Stevens Honor System.

Project Description:

In this project, we worked on predicting whether a company would go bankrupt based on the financial data given. Since the dataset was highly imbalanced, the logistic regression model didn't perform well and failed to make accurate bankruptcy predictions.. To improve the results, we first reduced the number of features using PCA and grouped similar companies using k means clustering. Then, for each subgroup, we built stacking models with multiple non-parametric base models. Our goal was to make the predictions more accurate while keeping the models efficient and easy to manage.

2. Company Characterization for each sub group(Section 3.2):

Number of subgroups: 5

Sub Group 1:

Number of y = 0 and y = 1

Number of Samples = 1148

Number of Non Bankrupt Companies (y=0) = 1107

Number of Bankrupt Companies (y=1) = 41

Bankruptcy rate = 3.57%

Selected Features: PC1, PC2, PC3, PC5, PC7, PC11, PC19, PC20, PC23, PC24 (10 Features)

Summary of Characteristics :

- Companies in Cluster 1 show **moderate variability** across principal components PC1 to PC26.
- The **means** of most principal components are **close to zero**, but a few components (such as PC1 and PC3) have **more negative means**, suggesting that companies differ in their underlying financial patterns captured by those components.
- The **standard deviations** of principal components typically range from **1.5 to 2.0**, indicating **moderate spread** of financial behaviors within the cluster.

Summary of Distribution and Properties:

- **PC1 and PC3**: Both show **negative mean values** with moderate dispersion (std ~1.6–2.0), suggesting that along these dimensions, companies tend to cluster toward lower scores.
- **PC2, PC5, and PC7**: These components have **positive means**, indicating higher values along these particular PCA directions among the majority of companies.
- **Later principal components (PC19, PC20, PC23, PC24)**: Have **means near zero** and **lower variability**, suggesting that variation along these dimensions is less pronounced.

Overall spread:

- No severe skewness detected across PCs.

- Distributions appear relatively normal with some variability around the means.

Conclusion:

- Companies in Subgroup 1 display **moderate stability across principal components**, with **some dispersion along key PCA directions** but no extreme outliers or clustering anomalies.
- The bankruptcy rate is relatively low at **3.57%**, indicating overall financial health, though some variability exists across underlying PCA dimensions.

Sub Group 2:

Number of y = 0 and y = 1:

Number of Non Bankrupt Companies (y=0) = 1544

Number of Bankrupt Companies (y=1) = 3

Total companies: 1547

Summary of Characteristics:

- PC1 (which heavily loads on "ROA before interest and depreciation") is highly positive. Companies generally show good return on assets before interest and depreciation.
- PC11, PC17, PC23, and PC12 also have positive scores. These likely relate to tax rates, profit growth rates and net value per share thereby showing a stable profitability.
- PC3, PC6, PC5, PC4 are strongly negative. These PCs are related to Operating Margins, Non-industry income, and Interest Rates. This indicates that companies in Cluster 2 might have lower operating margins and higher non-industry costs.

Summary of Distribution:

- Companies in Cluster 2 seem moderately profitable but might struggle with efficiency or external costs.

Bankruptcy Rate for Cluster 2:

- We can see that the bankruptcy rate for cluster 2 is very low at 0.19% (only 3 out of 1547 companies). This indicates that companies in this cluster are generally financially stable.

Properties of Companies in Cluster 2:

- Cluster 2 primarily consists of financially stable companies with strong returns on assets and consistent growth indicators.
- However, they exhibit slightly lower operating margins and face moderate external non-industry costs.
- The bankruptcy rate in this cluster is very low, indicating overall healthy financial conditions.

Sub Group 3:

Number of y = 0 and y = 1

Number of Non Bankrupt Companies (y=0) = 994

Number of Bankrupt Companies (y=1) = 11

Summary of Characteristics:

Companies in Subgroup 3 show strong clustering around higher PC1 and PC2 values, suggesting better profitability and operational margins than random distribution whereas, PC3, PC5, and PC7 exhibit slight negative means, indicating moderate risks tied to operational costs and asset management.

Subgroup 3 companies have medium-level variability in financial behavior across PCs but maintain positive trends in core financial indicators

Summary of Distribution:

Most financial features are normally distributed, meaning most companies are grouped around the average, with a few companies on the lower side. PC1 and PC2 have positive means and relatively low standard deviations, indicating they are financially strong.

Properties of Companies:

Companies in Subgroup 3 are generally financially stable, includes about 11 bankruptcies, which is about 1.09% bankruptcy rate. Financial health indicators such as ROA and operating margins are positive. Overall, companies in this subgroup require moderate monitoring, especially those with lower scores on PC5 and PC7.

Sub Group 4:

Number of y = 0 and y = 1

Number of Non Bankrupt Companies (y=0) = 992

Number of Bankrupt Companies (y=1) = 143

Summary of Characteristics:

PC1 (which heavily loads on "ROA before interest and depreciation") shows lower scores for bankrupt companies.

- 1)Companies with lower PC1 values are more likely to be financially distressed.
- 2)PC1 captures profitability strength — lower ROA is strongly associated with higher bankruptcy risk.

PC4, PC5, and PC6 (related to Operating Margins and Realized Sales Gross Margin) show important separation.

- Higher operating profitability is a positive signal for company survival.

PC13, PC14, and PC15 (related to Expense Rate, R&D Expense Rate, and Cash Flow Rate) show little separation.

- These features may not be strong individual indicators for bankruptcy within Cluster 4.

PC22, PC24, and PC25 (related to per-share profit and pre-tax earnings) show moderate contributions.

- Healthy per-share profit metrics help non-bankrupt companies stand apart.

Summary of Distribution:

Companies in Cluster 4 show moderate financial health overall but display early warning signs of financial distress, especially through lower ROA and weaker operating margins.

Bankruptcy Rate for Cluster 4:

The bankruptcy rate for Cluster 4 is moderate at approximately 12.6% (143 bankrupt companies out of 1135 total companies).

- This is higher than extremely stable clusters like Cluster 2 but is not extremely alarming.

- Indicates that proactive monitoring is needed.

Properties of Companies in Cluster 4:

Cluster 4 primarily consists of companies with moderate financial stability, where profitability ratios like ROA and Operating Margins play a crucial role in distinguishing financially healthy firms from distressed ones.While the majority of firms are stable, a noticeable minority show signs of financial weakness, particularly those with lower returns on assets and weaker operating profitability.Key financial ratios must be carefully monitored to differentiate distressed companies from healthy ones within this subgroup**.**

Sub Group 5:

Number of y = 0 and y = 1

Number of Non Bankrupt Companies (y=0) = 972

Number of Bankrupt Companies (y=1) = 0

Total Number of Companies = 972

Summary of Characteristics:

- PC1 has a very high mean. This means that companies in Cluster 5 show very strong performance in core financial metrics like Return on Assets (ROA) and profit before interest and depreciation.
- PC4, PC3, PC5, and PC6 are also strongly positive. These are likely tied to operating

profit rate, gross margins, and non-industry earnings — signs of operational and financial health.

- PC2, PC11, and PC17 are the most negative. This suggests that slightly weaker performance in tax efficiency, EPS consistency, or long-term growth, but not enough to impact bankruptcy.

Summary of Distribution:

The Bankruptcy rate for this cluster is 0%. This means that Cluster 5 has all the companies that are Financially stable.

Properties of Companies:

- Cluster 5 includes 972 companies, all of which are financially stable with no bankruptcies. The cluster exhibits high scores in key components such as ROA, operating profits, and gross margins, indicating a highly profitable and operationally efficient group.
- Although a few PCs like PC2 and PC11 show mild negative trends, they are not substantial enough to affect the group's overall strength.
- The financial health of companies in this cluster is very strong.

---

3. Building Training Models (Section 3.3):

Table 3:

Subgroup ID	Name of Student	Average accuracy score base models [TT(TF)]	accuracy score Meta model [TT(TF)]	$N_{features}$
1	Elina Rezaeian	0.99[40(1)]	0.90 [37(4)]	10
2	Ishaan Chawathe	0.667 [6(3)]	1.00 [3(0)]	20
3	Hargun Sawhney	1.0[11(0)]	0.909 [10(1)]	8
4	Ottikunta Sahith	0.895[128(15)]	0.853[122(21)]	9
5	Constant	1.00[972(0)]	1.00[972(0)]	0
Team		0.91 [1157(19)]	**0.93** [1144(26)]	**9.4**

4. Generalization:

In the generalization process, we started by preprocessing the test data through feature scaling and dimensionality reduction using PCA. Next, we assigned each company to a cluster using our pre-trained cluster classifier. For each cluster, we applied its specific trained model to predict whether a company would go bankrupt. Companies in Cluster 5, which had no bankruptcies during training, were directly predicted as non-bankrupt. All individual predictions were then compiled into a single file in the required submission format. The Table can be found in 5_Generalization.csv file.

5. Conclusion:

- In this project, we built a custom machine learning model using stacking to predict company bankruptcies. We started with 5807 companies and 95 features, and after cleaning the data and reducing it to 26 key components using PCA, we grouped the companies into five clusters based on financial similarity.
- Each cluster had its own model, allowing us to better handle different risk profiles. Cluster 2 had a very low bankruptcy rate of 0.19%, and no bankruptcies were observed in Cluster 5l. Clusters 1, 3, and 4 had slightly higher rates of 3.57%, 1.09%, and 12.60% respectively. Cluster 4 has the most bankruptcies overall (143 out of 1135). This setup helped us build models that were more focused and accurate.
- For the test phase, we transformed the test data like the train data using the same process. We cleaned the data, applied PCA, assigned each company to a cluster, and used the relevant model to make predictions. Cluster 5 companies were automatically marked as non-bankrupt. In the end, our predictions were saved in a separate .csv file. In this way, the approach helped us to improve bankruptcy detection and overcome the limitations of standard models like logistic regression which had nearly zero recall on bankrupt companies.

6. Video Presentation :

- Please find the the video attached in the zipped folder.

Thank you