IE7275 Fall 2021 Term Project Report
Mahrukh Malik, Sonali Godade, Sumanth Wannur

Bankruptcy Prediction Using Supervised Learning Methods

1. Team Responsibility

Exploratory Data Analysis (EDA)

1. Sonali Godade

Data Preprocessing

1. Mahrukh Malik

Data Mining Methods

1. Mahrukh Malik
2. Sumanth Wannur
3. Sonali Godade

Outcomes Evaluation

1. Sumanth Wannur

Final Report Generation

1. Sonali Godade
2. Sumanth Wannur
3. Mahrukh Malik

Final Presentation

1. Sonali Godade
2. Sumanth Wannur
3. Mahrukh Malik

2. Abstract

Early detection of a company's financial distress is critical. Recall the financial crisis of 2008; many stakeholders were caught unprepared. A select few experts were able to recognize the early warning signs of an approaching calamity, but the majority did not have the tools to systematically identify these signals or the ability to decipher intricate and complex patterns in data. Thanks to developments in data analytics, we now have the chance to alter this narrative. The challenging aspect of the task was dealing with class imbalance due to the rarity of the bankruptcy event. In this work, to understand the impact of different inputs and their respective results with a suitable model, we derived seven datasets from the main source. It was mainly divided into two parts: one is financial metrics as given in the main dataset, and the other is the famous Ohlson's financial ratios. Based on the dataset, we selected suitable models and derived some insights, shedding light on important aspects of the project. We determined that panel data and datasets with lagged values yield more accurate results, with the GBM model performing the best on both types of data. Additionally, financial metrics offer deeper insights into the financial health of a company, resulting in more accurate predictions.

3. Introduction

Understanding the story that the numbers convey is just as important to bankruptcy prediction as the numbers themselves. Leveraging data analytics to predict the financial condition of business organizations comes with a twofold benefit for consultancies and financial experts. First, seeing when a company is nearing a financial crisis is an opportunity to intervene and provide data-driven solutions. Second, with a reliable bankruptcy prediction model, consultancies can approach potential clients proactively, bringing value when it's needed most. The ability to predict bankruptcy can also be utilized by banks to assess the creditworthiness of their corporate clients and by financial regulatory bodies to assess the economic health of the industries and make timely interventions.

Bankruptcy is a legal proceeding initiated when a business is unable to repay outstanding debts or obligations. In our analysis, we specifically considered Chapter 7 bankruptcy, where the company ceases all operations and completely goes out of business. A trustee is appointed to liquidate the company's assets, and the proceeds are used to pay off debts.

Predicting future financial crises is not just an advantage, but a necessity in today's data-driven business landscape. Organizations operate in conditions full of uncertainties, from shifting consumer needs to unforeseen global trends and events. Bankruptcy is one of the most debilitating challenges a company can encounter. It has repercussions for investors, employees, and the market at large, in addition to the company. Therefore, timely prediction of an impending financial distress is critical not only for the companies but also for consultancies that provide services and solutions to companies in crises. Finding at-risk companies is about more than just averting a financial crisis. This forecast presents consultancies with a window of opportunity to reach out to prospective clients and provide proactive solutions, thereby solidifying their position as proactive market participants and reliable advisors.

4. Problem Definition

Our challenge is to develop a reliable predictive tool that tracks financial health indicators by using machine learning. This model should find patterns in large datasets that could be missed in more conventional analysis. The goals are to:

- Utilize advanced algorithms to identify early indicators of financial crisis.
- Provide consultants with actionable data-driven insights so they can provide strategic solutions.

The sheer volume of financial data and the thousands of companies operating in a complex business environment means that traditional methods of risk assessment might not be sufficient. Because of the intricacy of the financial indicators and how they interact, advance data-driven techniques are needed to process and analyze the data at scale and find patterns and correlations that may be imperceptible to traditional methods

5. Data Description

We obtained our data from Standard & Poor's Compustat through the WRDS platform. WRDS is a premier data platform offering a suite of financial and economic data to academics and professionals. Compustat is a database of financial, statistical, and market information on active and inactive global companies throughout the world. The initial data extract comprised of 974 variables and 4 million records, containing data for all the publicly listed companies globally. Since the data on Compustat comes from a variety of different sources, such as companies' financial statements, market data, auditor data etc., the data contained a lot of null values and was inconsistent over the years for some organizations.

We narrowed our focus to American countries to ensure regulatory consistency as companies across the globe are governed by different financial reporting standards and therefore, comparing companies using different standards is not a straightforward task. Secondly, we wanted to ensure economic homogeneity and market conditions. Additionally, we wanted to focus on a specific legal framework as bankruptcy laws differ significantly from country to country.

From the original 974 variables, we extracted 18 financial indicators based on their subsequent impact on a company's financial health. After removing null values, we were left with 8,971 companies—8,362 alive and 609 failed—resulting in a total of 78,682 records spanning over 20 years from 1998 to 2018. The number of records varies each year, with 5,308 records for 1999 and 2,723 records for 2018.

Variables

The final variables we selected from the total 974 variables in the original data are as follows:

- Current Assets: Represents company's assets that are expected to be used up or converted into cash in the next year.

- COGS: Denotes company's total sum of expenses that are directly incurred due to sale of products and services.
- Depreciation and Amortization: Cumulative reduction in value of company’s tangible and intangible assets over time.
- Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA): Denotes a company's earnings before subtracting interest, taxes, amortization, and depreciation.
- Inventory: Total value of raw material and final goods in the inventory. This is a snapshot of a point in time.
- Net Income: Represents a company's final profit after deducting all its expenses from its revenue.
- Total Receivables: The total amount owed to a business for availed goods or services that customers have yet to pay.
- Market Value: Total market value of shares of publicly listed companies at the time of valuation.
- Net Sales: The resulting sum after subtracting returns, allowances, and discounts from a company's gross sales.
- Total Assets: Represents all of the company’s valuable assets.
- Total Long-term Debt: Pertains to all loans and debts owed by a company that have a maturity date beyond one year.
- EBIT: Represents a company’s earnings before interest and taxes deduction.
- Gross Profit: The residual profit after subtracting the costs related directly to the production and sale of goods or services.
- Total Current Liabilities: the sum of all debts that must be paid before the end of the fiscal year. This includes accounts payable, bonds due, unpaid wages and salaries, and other similar obligations.
- Retained Earnings: Profit remaining after deducting all operating expenses, taxes, and shareholder dividends.
- Total Revenue: The sum of all sales before any costs are deducted. Investment income like interest and dividends on investments could be included.
- Total Liabilities: All of a company's debts and financial obligations to third parties.
- Total Operating Expenses: The total costs incurred during regular business activities.

Exploratory Data Analysis

The distribution of the target variable reveals a notable class imbalance, with 73,462 records classified as "alive" and 5,220 records classified as "failed" (see figure 5.1). This imbalance poses a challenge in model development, as the majority of businesses are labeled as "alive," potentially leading to biased predictions. Addressing this imbalance during the model-building process is crucial to ensure accurate and reliable results.

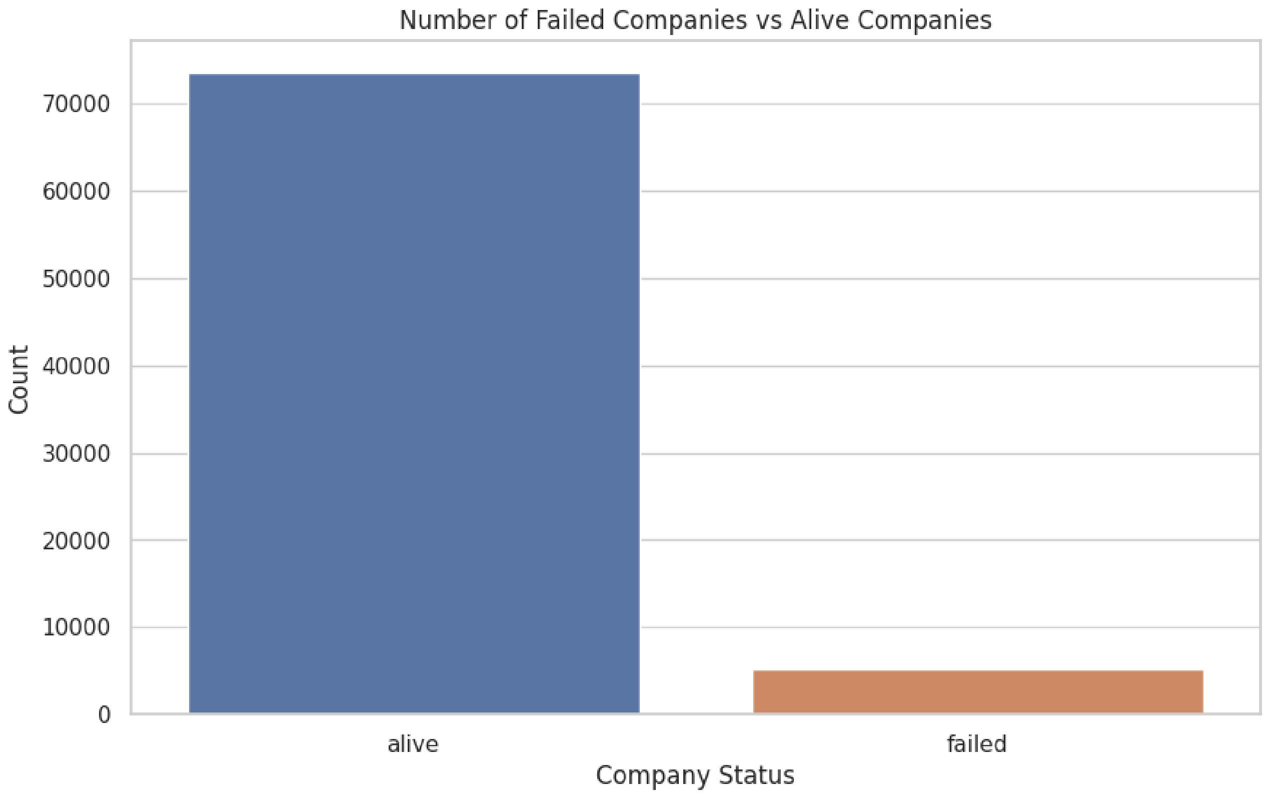


Figure 5.1

We also observed data overlap across all features, for both classes (see figure 5.2). For example, the range of current assets for failed companies (Status 1) falls entirely within the range of current assets for alive companies (Status 0). Specifically, the overlap occurs in the lower range of current assets. Since all the data points for failed companies (Status 1) are also covered by the spread of points for alive companies (Status 0), the overlap is in the entire range of current assets values represented by failed companies.

This may be attributed to the ground reality that even though a company is not going bankrupt, it may still be undergoing financial distress or events like leveraged buyouts, restructuring, or significant shifts in market conditions that affect its financial metrics similarly to a failing company.

Financial distress can significantly reduce a company's liquidity and current assets, even if it does not lead to bankruptcy. For instance, a company engaged in a leveraged buyout might take on substantial debt, reducing its net current assets. Such scenarios can cause the financial profiles of distressed but still 'alive' companies to resemble those of companies that are actually going bankrupt, leading to the observed overlap in the distribution of current assets between the two groups.

This overlap suggested that for any given financial metric, there were both alive and failed companies with similar values, making it challenging to distinguish between the two. Consequently, this complicated the task of building a predictive model, as it would require more sophisticated techniques or additional features to effectively separate the classes and accurately predict the likelihood of a company going bankrupt.

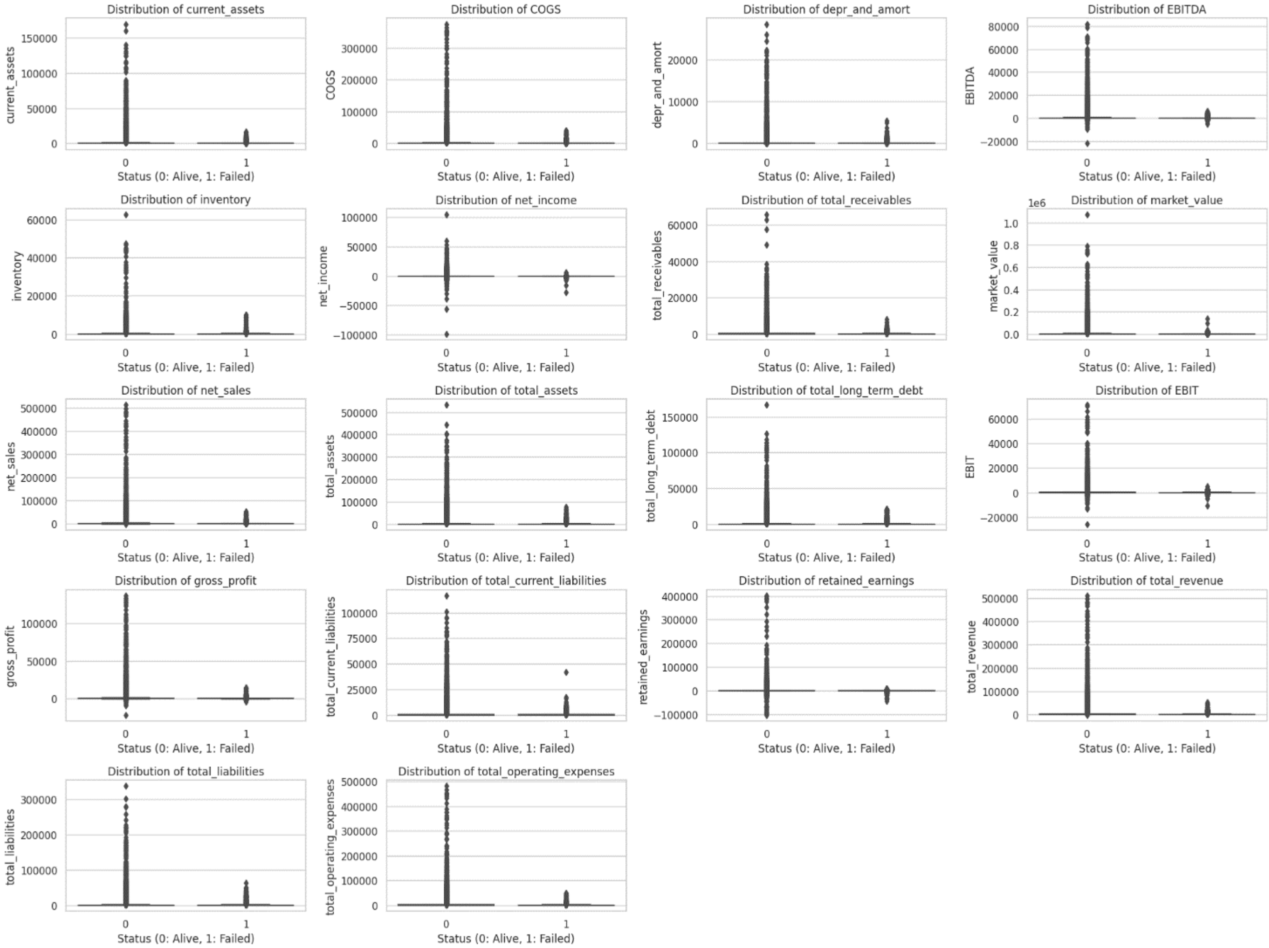


Figure 5.2

Figure 5.3 indicates a notable decrease in the number of company bankruptcies after the 2007 financial crisis. This decline could be due to stricter financial regulations and more cautious business practices. As a result, bankruptcies have become more infrequent, exacerbating the class imbalance in the data, particularly in the years following the crisis. This makes it challenging for predictive models to learn from a dwindling number of failure cases, potentially affecting their accuracy in forecasting future bankruptcies.

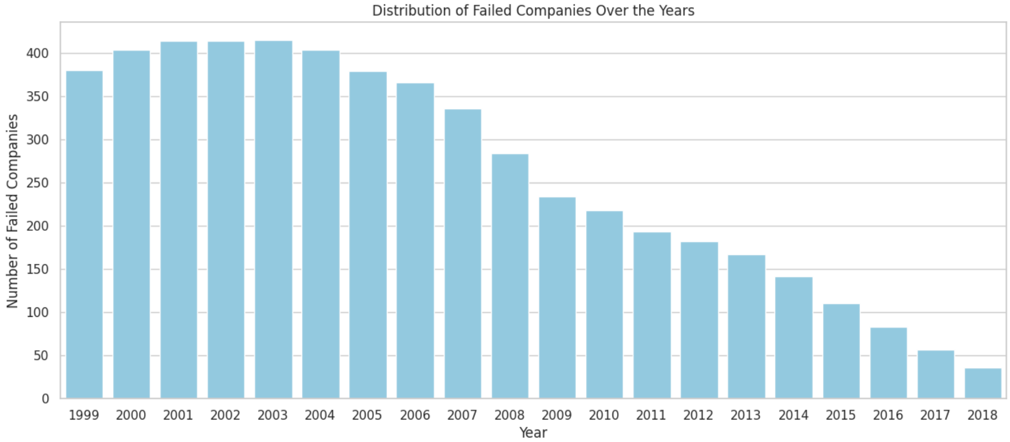


Figure 5.3

We also had substantial of multicollinearity in our predictor variables. The matrix in figure 5.4 demonstrates strong positive correlations between many of the financial variables. Such high correlations are common in financial data due to the interconnected nature of financial statements where certain variables are intrinsically related (like sales and revenue, or assets and liabilities). The presence of strong correlations close to 1 suggest multicollinearity, which can be a concern when building regression models.

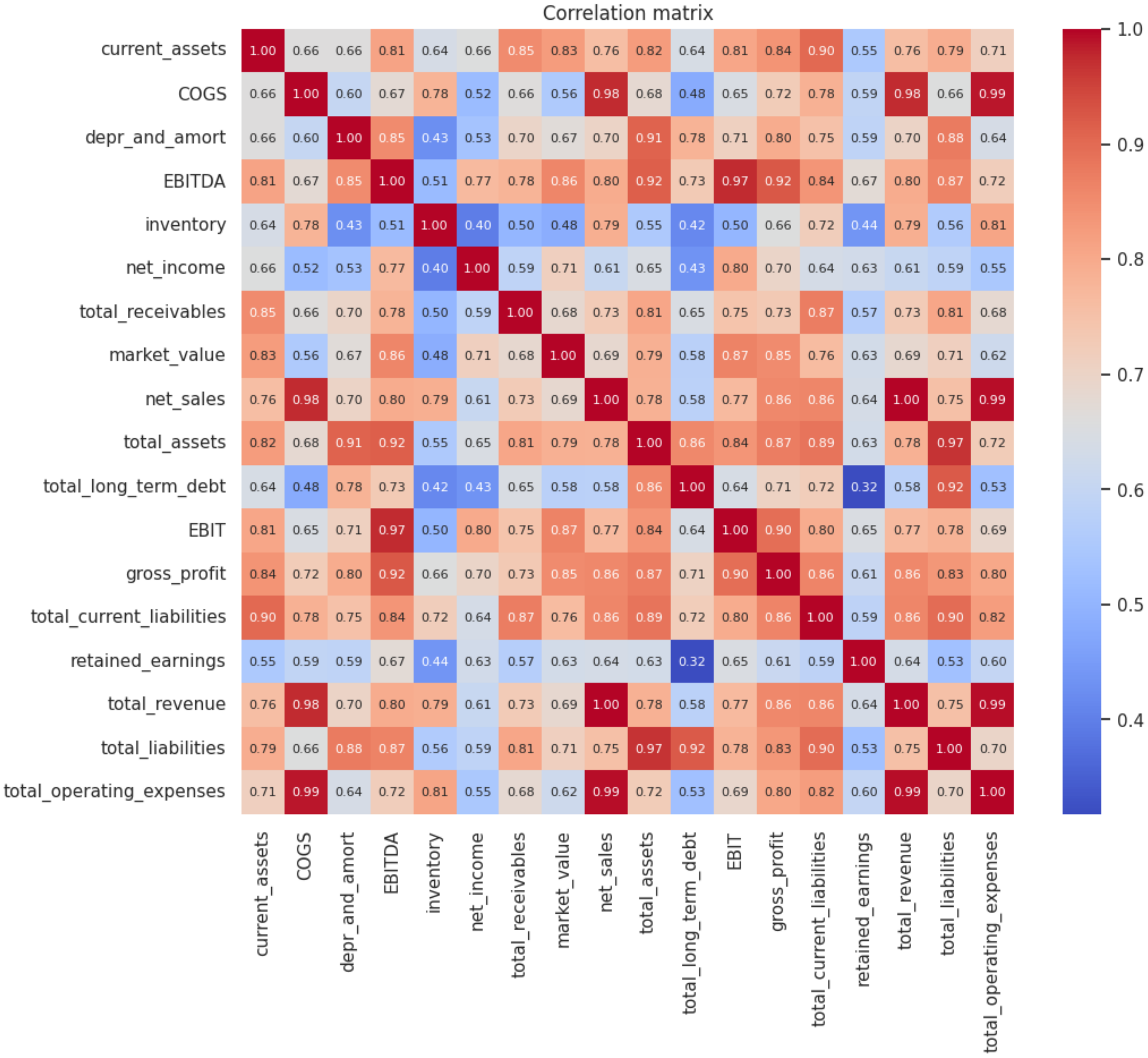


Figure 5.4

6. Proposed Methods

Panel data is a dataset that contains observations on a set of entities—such as individuals, organizations, or countries, across time. Our data qualifies as panel data because it tracks various financial metrics for a set of companies over several years.

Standard regression models typically assume that each observation is independent of the others. However, in panel data, observations for the same entity are likely to be correlated over time, violating this assumption. This is why panel data cannot be directly used in classical regression models without adjustments to allow for cross-sectional analysis (where each sample is considered independant)

To analyze such panel data effectively, we employ three different techniques, each offering a unique lens:

- **Using Lagged Values:** By analyzing the data in its original form, we can observe how each company's metrics evolve over time, thus maintaining the integrity of the temporal information.
- **Handling as Panel Data:** It leverages the data's longitudinal structure to examine temporal trends and the evolution of financial health over time.
- **Cross-sectional Snapshots:** We take a specific time slice of data for each company, effectively treating each company's selected observation as independent. Simplifying the data to one year observation per company allows us to focus on specific years, particularly important in contrasting the conditions between companies that failed and those that did not at a particular time.

Consequently, we conducted experiments on seven datasets derived from the main source. Our data is divided into two main parts based on the type of variables used to derive the results. The first part involves financial metrics where we have taken variables directly from the main source. The second part involves Ohslon’s financial ratios; famous ratios that have been used since 1970s to predict bankruptcy. Furthermore, both datasets are further divided into three sub-datasets.

- **Dataset 1:** This dataset contains lagged variables. We have dataset 1 for both, financial metrics as predictors as well as financial ratios as predictors.
- **Dataset 2:** This dataset contains the original panel data and is also available for both financial metrics as well as financial ratios.
- **Dataset 3:** This data contains cross-sectional snapshots of data for both financial metrics and financial ratios of companies.

Additionally, financial metrics data is modeled by applying PCA without any form of feature downsizing, serving as the seventh dataset.

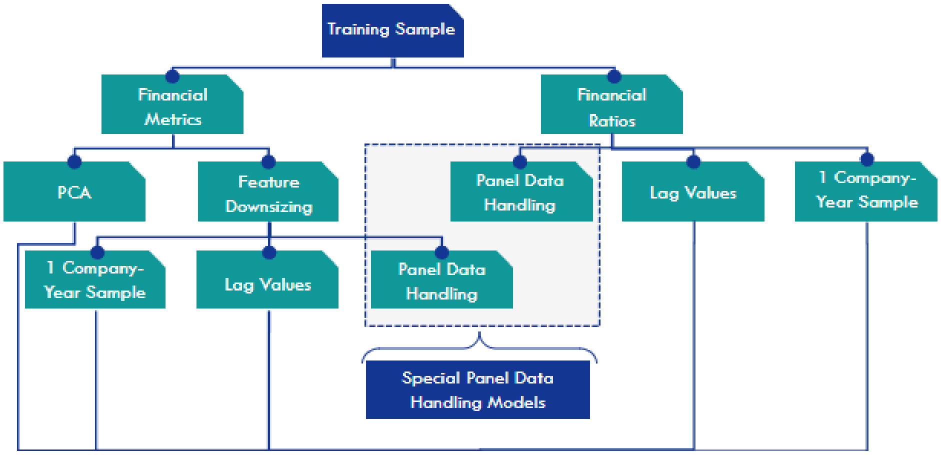


Figure 6.1

Feature Engineering

As mentioned, we employed three different kinds of feature techniques.

i. Feature Downsizing for Datasets with Financial Metrics as Predictors:

Given that the financial metrics had a lot of multicollinearity, we used feature engineering to minimize the multicollinearity in our datasets:

- Net_sales and total_revenue are effectively the same and thus exhibit perfect correlation (1.00). To reduce multicollinearity, we removed total_revenue from our analysis.
- Net sales and COGS are also closely linked. Since gross profit is net sales minus COGS, it reflects the necessary information from both. Moreover, as COGS is included in total operating expenses, its removal along with net sales does not result in information loss. To reduce multicollinearity, we excluded net sales and COGS, retaining gross profit and total operating expenses for a clearer analysis.
- "EBITDA" and "EBIT" show a high degree of correlation since "depr_and_amort" is the only differentiator, defined as EBITDA - EBIT = depr_and_amort. Therefore, to lower multicollinearity, EBITDA was omitted from the dataset as that information is present in EBIT and depr_and_amort.
- Since "gross_profit" is EBIT + depr_and_amort + total_operating_expenses, it is redundant and was removed from the dataset.
- "total_assets" and "total_liabilities" are highly correlated. Since equity = assets - liabilities, this correlation might reflect similar equity ratios across companies. It's not clear if either should be removed, hence we combined their information into a single variable by calculating the ratio of assets to liabilities.

After applying the aforementioned feature selection, we were able to reduce multicollinearity significantly as can be observed from the updated correlation matrix in figure 6.2.

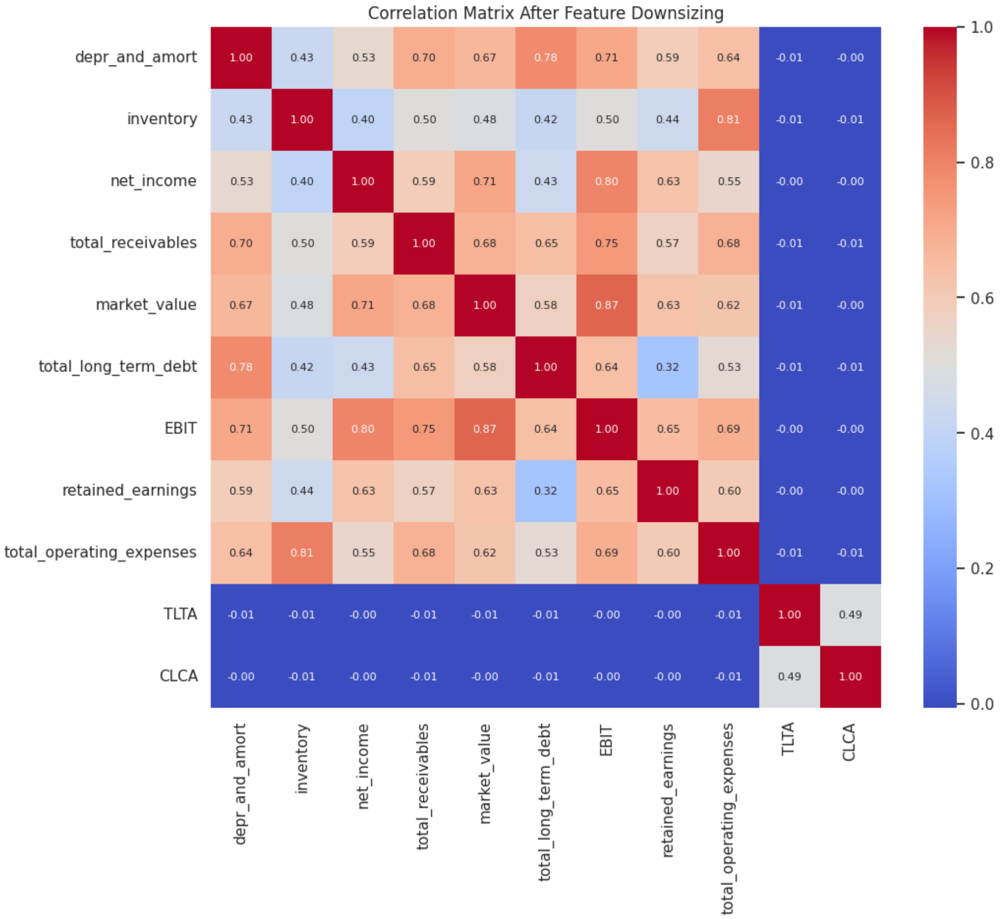
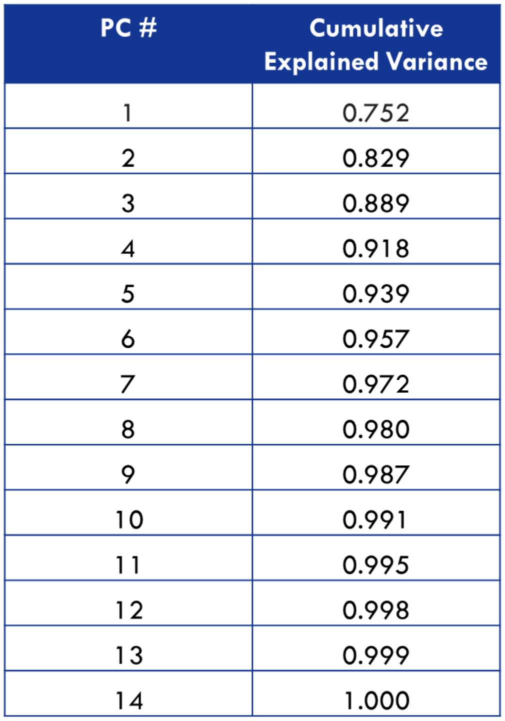


Figure 6.2

ii. Principal Component Analysis

We also conducted PCA on the original dataset to evaluate if reducing multicollinearity through PCA can improve results for regression models. Given the cumulative explained variance table in Figure 6.3, we decided to retain 7 principal components that explain 97% of the variance in the data:



PC #	Cumulative Explained Variance
1	0.752
2	0.829
3	0.889
4	0.918
5	0.939
6	0.957
7	0.972
8	0.980
9	0.987
10	0.991
11	0.995
12	0.998
13	0.999
14	1.000

Figure 6.3

iii. Computing Financial Ratios

Altman's financial ratios are a set of five different financial metrics that were combined by Edward Altman into a single score known as the Altman Z-score. This score is used to predict the likelihood of a business going bankrupt within a two-year period. The Z-score model was developed in 1968 and it is particularly well-regarded for its accuracy in predicting bankruptcy, showing an accuracy of 72% in predicting bankruptcy two years prior to the event according to some studies.

The key financial ratios used in the Alman model include:

- X1 (liquidity) = Current assets – Total liabilities / Total Assets
- X2 (profitability) = Retained earnings / Total Assets
- X3 (productivity) = EBIT / Total assets
- X4 (debt ratio) = Market value / Total long term debt
- X5 (asset turnover) = Net sales / Total assets

After calculation of financial ratios, the multicollinearity in the predictors was substantially less than the multicollinearity between financial metrics in the original data, as can be observed from the matrix in Figure 6.4.

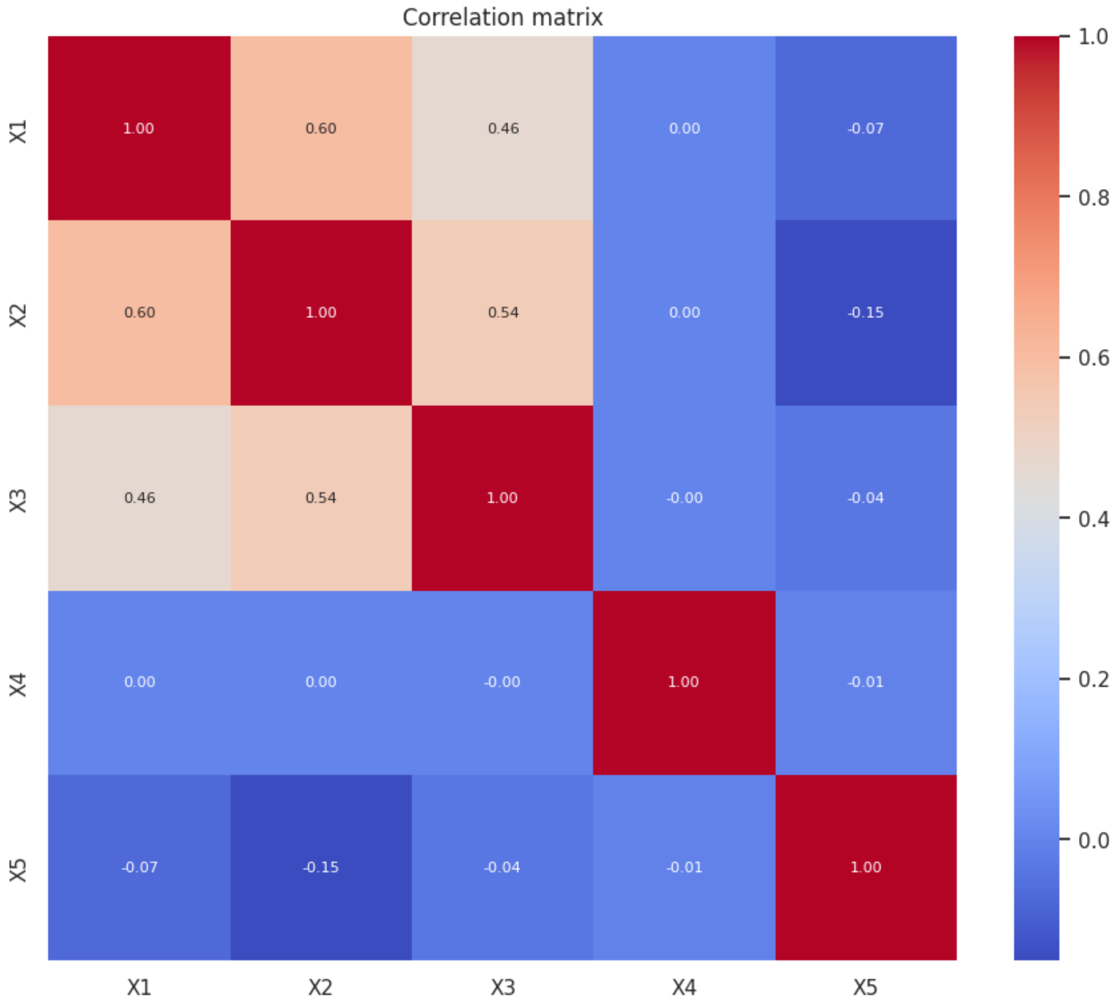


Figure 6.4

Model Selection

The goal is to determine the most effective data mining algorithms that can predict bankruptcies. One approach is to use simple classification techniques. Another method is to leverage the time series structure of the data as we have observed that the dataset is structured with multiple years of data for many companies. This allows us to track the financial health of a company over time. A company's likelihood of filing for bankruptcy increases, for instance, if its financial metrics continue to worsen for several years in a row.

As discussed, the classical regression models cannot be employed on panel data, therefore, we applied a separate set of models on dataset 2 (panel data) for both sets of predictions, i.e. financial metrics and financial ratios. Given below is the list of models we tested:

i. **Model selected for data without panel structure:**

a. **Logistic Regression (LR)**

This technique evaluates the relationship between one or more independent variables and the categorical dependent variable. In order to predict an event, it provides probabilities and categorizes them, making it appropriate for binary outcomes like bankruptcy.

b. **Random Forest (RF)**

Random forests can capture complex data structures and handle high dimensionality well, making them apt for financial predictions

c. K-Nearest Neighbor Classifier (KNN)
KNN uses neighbor classification to classify data points. If our data contains distinct clusters of companies that are in bankruptcy and those that are not, then it might be useful.

d. Support Vector Machines (SVM)
SVM is a powerful machine learning algorithm for classification and regression, excelling in scenarios requiring clear class boundaries or benefiting from mapping data to higher dimensions. SVMs identify optimal hyperplanes for class separation in complex datasets, proving valuable for non-linearly separable data points in their original feature space.

e. Generalized Estimating Equations (GEE)
Gradient Boosting is an ensemble machine learning technique that builds and optimizes predictive models in a step-by-step fashion. It improves predictive accuracy by sequentially adding weak models, typically decision trees, to correct the errors of the existing model.

ii. Model for Panel Data Structure:

a. Generalized Estimating Equation:
They are particularly useful for analyzing panel data because they provide a flexible and robust way to handle correlated observations within the same cluster, which is common in datasets that span over time. Additionally, GEE can handle unbalanced data, which is often the case with panel datasets.

b. Long Short-Term Memory networks:
LSTMs for capturing dependencies and patterns in sequential data. LSTMs come with a memory cell that can capture long-term dependencies in the data, and they're versatile enough to handle variable-length sequences, making them a good fit for unbalanced panel datasets.

c. Gradient Boosting Machines:
GBMs automatically assess variable importance. This helps us identify which features are more influential in predicting the outcome variable over time. GBMs are valuable for understanding the dynamics of panel data, as they are robust to outliers, handle non-linearity, and capture temporal trends while adapting to changing patterns over time.

Hyperparameters Tuning

To enhance the performance of our models, we focused on fine-tuning the hyperparameters of those that demonstrate superior performance on validation data. For this purpose, we have employed the Optuna library. Optuna is adept at streamlining the hyperparameter optimization process by facilitating efficient parallelization. This approach allows us to systematically explore a range of hyperparameter configurations to identify the most effective settings for our models.

Performance Evaluation

The following metrics were taken into consideration when evaluating the models' performance:

- **Accuracy:** A broad indicator of how frequently the model makes accurate predictions.
- **Sensitivity (Recall):** This metric, which quantifies how frequently the model detects actual bankruptcy cases, is fundamental for our use case.

Evaluation Strategy:

- **Data Splitting** was used as the evaluation strategy to separate the dataset into training and test sets. The test set was used to assess performance, and the training set was used to train the models.
- **Balancing data:** To balance the data, under sampling method was employed.

- **Post-Evaluation:** We ranked the models after training and evaluation according to their sensitivity and accuracy metrics.

In conclusion, in order to determine which models are best suited for bankruptcy prediction, the process involved rigorous training, evaluation, and fine-tuning. The aforementioned approaches have been selected due to their theoretical utility in classification problems, and their performance is empirically verified on our dataset.

7. Experimental Results

Figure 7.1 shows recall and accuracy scores of our models:

Data type	Data set	Model	Recall Score	Accuracy
Panel Data	Financial Metrics	GEE	84%	68%
	Financial Ratios		2%	97%
	Financial Metrics	LSTM	6%	91%
	Financial Ratios		1%	89%
	Financial Metrics	GBM	79%	71%
	Financial Ratios		84%	63%
Lagged data	Financial Metrics	LR	87%	32%
	Financial Ratios		36%	73%
	Financial Metrics	RF	67%	71%
	Financial Ratios		74%	69%
	Financial Metrics	KNN	76%	64%
	Financial Ratios		65%	59%
	Financial Metrics	SVM	98%	22%
	Financial Ratios		71%	58%
	Financial Metrics	GBM	70%	67%
	Financial Ratios		78%	65%
Single year data	Financial Metrics	LR	83%	44%
	Financial Ratios		20%	66%
	Financial Metrics	RF	70%	58%
	Financial Ratios		78%	68%
	Financial Metrics	KNN	80%	43%
	Financial Ratios		80%	50%
	Financial Metrics	SVM	87%	30%
	Financial Ratios		100%	4%
	Financial Metrics	GBM	54%	76%
	Financial Ratios		78%	62%
PCA		LR	90%	33%
		RF	76%	69%
		KNN	63%	79%
		SVM	98%	19%
		GBM	82%	60%

8. Conclusion (for final report)

When selecting the best models based on a combination of recall and accuracy, we need to consider the trade-offs between these two metrics. Higher recall means that the model correctly identifies more of the bankruptcy cases, which is important in our use case. However, accuracy is also important as it reflects the overall correctness of the model across all cases.

Here are some models that stand out based on their balance of recall and accuracy:

- **Panel Data - Financial Ratios with GBM:** 84% Recall and 63% Accuracy. This model has the highest recall of all the models for financial ratios and reasonably high accuracy.

- **Panel Data - Financial Metrics with GBM:** 79% Recall and 71% Accuracy. This model has the best combination of recall and accuracy.

- **Lagged Data - Financial Metrics with GBM:** 70% Recall and 67% Accuracy. While the recall is not the highest, this model has a good balance between recall and accuracy.

- **Lagged Data - Financial Ratios with GBM:** 78% Recall and 65% Accuracy. The recall is fairly high but at the expense of some accuracy.

- **Single Year Data - Financial Metrics with RF:** Random Forest Classifier did well on cross-sectional snapshots of data, with 78% Recall and 68% Accuracy. While the recall and accuracy are not the highest, the combination outperforms several other models.

- **PCA - KNN:** 63% Recall and 79% Accuracy. This model has a good balance, with a relatively high recall and the highest accuracy among the PCA models.

- **PCA - RF:** 76% Recall and 69% Accuracy. This model also has a good balance of recall and accuracy in the PCA category.

It's important to note that the choice of the model should be aligned with the specific business or research objective. If missing a positive case is very costly, a model with a higher recall might be preferred, even at the expense of accuracy. Conversely, if overall correctness across all predictions is more crucial, accuracy would be emphasized.

In summary, the evaluation of machine learning models for financial data analysis reveals that GBM models yield a high recall, particularly for Panel Data, suggesting their strength in identifying positive cases. For Lagged Data, GBM offers a balanced performance. In Single Year Data, the RF model stands out for its combined recall and accuracy. When using PCA, both KNN and RF show a commendable balance between recall and accuracy, with KNN being slightly more accurate and RF having a higher recall. Ultimately, the choice between these models should be based on the specific need for either recall or accuracy in the application at hand. These models are highlighted for their strong performance and provide a robust foundation for further model refinement.

References

1. https://sites.bu.edu/qm222projectcourse/files/2014/08/compustat_users_guide-2003.pdf
2. https://pages.stern.nyu.edu/~ealtman/Zscores.pdf
3. https://www.sciencedirect.com/topics/computer-science/bankruptcy-filing
4. https://www.sciencedirect.com/science/article/abs/pii/B978012370477150027X?fr=RR-2&ref=pdf_download&rr=826a11957aa13b81
5. https://www.sciencedirect.com/science/article/abs/pii/S0957417423019206