AI351 - Machine Learning Lab

Taiwanese Bankruptcy Statistics

Ali Hasan Khan
Artificial Intelligence(FCSE)
Ghulam Ishaq Khan Institute of Engineering and Technology
Topi,Swabi, Pakistan
u2021079@giki.edu.pk

I. Introduction

The provided dataset is in CSV format and encompasses a comprehensive collection of financial metrics and indicators for various companies. It comprises 96 columns and 6819 rows, with each row corresponding to a distinct company and each column representing a different financial ratio or indicator. The first column, labeled ”Bankrupt?”, serves as a binary indicator, denoting whether a particular company faced bankruptcy or not.

These financial ratios and indicators encompass a wide range of critical financial parameters that are essential for evaluating the financial well-being and performance of companies. Among the various ratios included are Return on Assets (ROA), operating profit rate, tax rate, debt ratio, and many others.

The dataset is a valuable resource for conducting in-depth analyses aimed at understanding the factors contributing to a company’s financial stability or its susceptibility to bankruptcy. By examining the data, researchers, analysts, and data scientists can gain insights into the financial health and risk profiles of different companies. This information can be leveraged to identify patterns, trends, and relationships between financial indicators and bankruptcy outcomes.

II. Exploratory Data Analysis

Exploratory Data Analysis (EDA) on the Taiwanese bankruptcy dataset is a critical step in understanding the financial factors associated with corporate bankruptcies. With 96 columns and 6819 rows, this dataset offers a comprehensive view of financial indicators, making it a valuable resource for researchers and analysts. Notably, the dataset boasts impeccable data quality, containing no missing values, which ensures the integrity of the analysis.

During EDA, analysts can employ statistical summaries and visualizations to gain insights into the dataset. Descriptive statistics help in identifying trends and variations in financial ratios, while visualizations like histograms and scatter plots make patterns more accessible. By exploring correlations between different financial metrics and the ”Bankrupt?” indicator, analysts can uncover valuable relationships.

Overall, EDA on the Taiwanese bankruptcy dataset, thanks to its data completeness, serves as a foundation for developing predictive models and strategies to mitigate financial risks. This analysis aids in identifying key financial indicators that contribute to bankruptcy, offering valuable guidance for decision-makers in the realm of corporate finance and risk management.

III. T-Test Statistics

Applying a t-test statistic to the columns ’ROA(C) before interest and depreciation before interest’ and ’ROA(A) before interest and

In statistical hypothesis testing, the t-test assesses whether the means of two groups are significantly different from each other. In this context, it suggests that there is a significant discrepancy between the ’ROA(C) before interest and depreciation before interest’ and ’ROA(A) before interest and

The p-value of 0.0 further corroborates the statistical significance of this difference. A p-value of 0.0 implies that the probability of observing such a significant difference between these ratios by random chance is extremely low. Typically, in hypothesis testing, a p-value below a predetermined significance level (e.g., 0.05) indicates that you can reject the null hypothesis, which in this case might be that the two ratios are equal.

In practical terms, these results suggest that there is a statistically significant dissimilarity between the financial indicators represented by these two columns. Further analysis can help uncover the underlying reasons for this difference and its potential implications for financial decision-making or risk assessment.

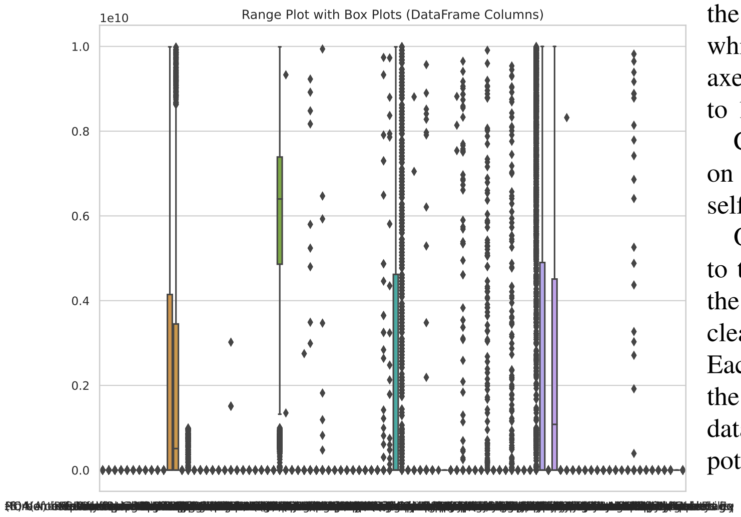
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

IV. Visualization

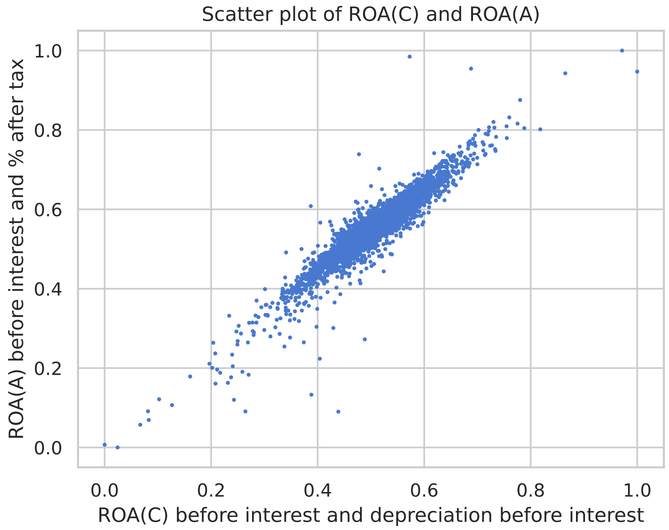
Using a vertical boxplot with Seaborn on a dataset containing 96 columns provides a powerful visualization tool for gaining insights into the distribution and variability of the data across multiple variables. A boxplot is a graphical representation that displays the summary statistics of a dataset, including the median, quartiles, and potential outliers. When oriented vertically, each column’s data is presented as a separate boxplot along the y-axis, making it easier to compare the distributions of numerous variables simultaneously.

This approach allows analysts to quickly identify central tendencies, such as the median or middle value, and assess the spread of data for each of the 96 columns. The length of the box represents the interquartile range (IQR), indicating the middle 50percentage of the data, while whiskers extend to show the range of the data within a certain threshold. Outliers, if present, are typically displayed as individual points beyond the whiskers.

The vertical boxplot in Seaborn is a valuable tool for identifying potential variations, skewness, and anomalies in a multi-column dataset. Analysts can utilize it to make informed decisions about data preprocessing, feature selection, or to gain preliminary insights into the dataset's overall distributional characteristics. By visually assessing the spread of data across numerous variables, it becomes easier to spot trends, variations, and potential data quality issues in a comprehensive manner.



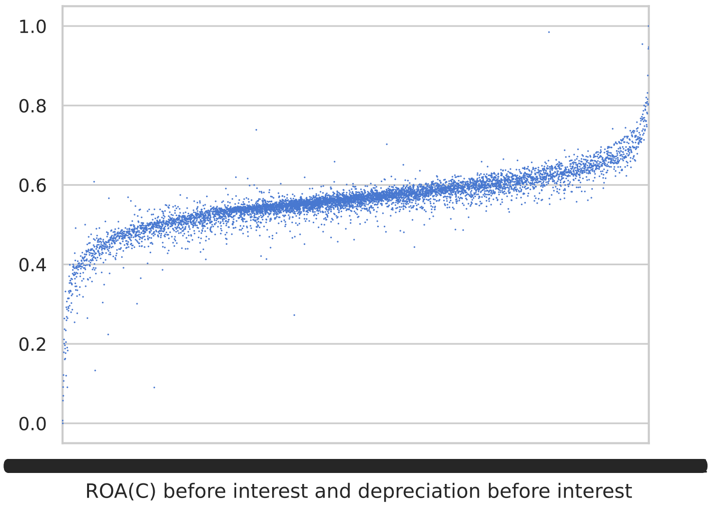
A scatter plot, featuring 'ROA(C) before interest and depreciation before interest' on the x-axis and 'ROA(A) before interest and percentage after tax' on the y-axis, offers a concise visual representation of the relationship between these financial metrics. Each data point on the plot represents a company, and the scatter of points reveals potential patterns or correlations. With compact points (s=2), it accommodates dense data well. Clear axis labels provide context, and the plot's title, 'Scatter plot of ROA(C) and ROA(A),' summarizes its purpose. Observing clustered or scattered points and their trends can quickly convey insights into the nature of the relationship between these two financial indicators, aiding further analysis and modeling.



Creating the Swarm Plot: sns.swarmplot() is used to create the swarm plot. In this plot, data points are represented as small individual points that do not overlap, making it easier to see the distribution of data. The data parameter specifies the DataFrame to use, and the x and y parameters determine which columns from the DataFrame to plot on the x and y axes, respectively. The s parameter sets the size of the points to 1, making them very small.

Customizing the Plot: ax.set(ylabel="") removes the label on the y-axis, which can be useful if the variable name is self-explanatory or if you want to save space on the plot.

Overall, this swarm plot offers an alternative visualization to the scatter plot discussed earlier. It allows you to examine the distribution of data points for each financial metric more clearly, especially when dealing with a large dataset like yours. Each point represents a company's financial performance for the two metrics, and the swarm plot helps identify how densely data points are clustered around different values, highlighting potential patterns or outliers in your financial data.


ROA(C) before interest and depreciation before interest