**Taiwan Economic Journal Analysis**

**1. Introduction**

The data was collected by the Taiwan Economic Journal for the years 1999 to 2009. The bankruptcy of the company was defined based on the corporate regulations of the Taiwan Stock Exchange.  
The goal is to be able to accurately predict the future bankruptcy of companies. The dataset contains 95 input and 1 output features. This is a classification task as it tries to predict whether the company will go into bankruptcy (label 1) or not (label 0), or the system is asked to specify which of the k categories an input belongs to.

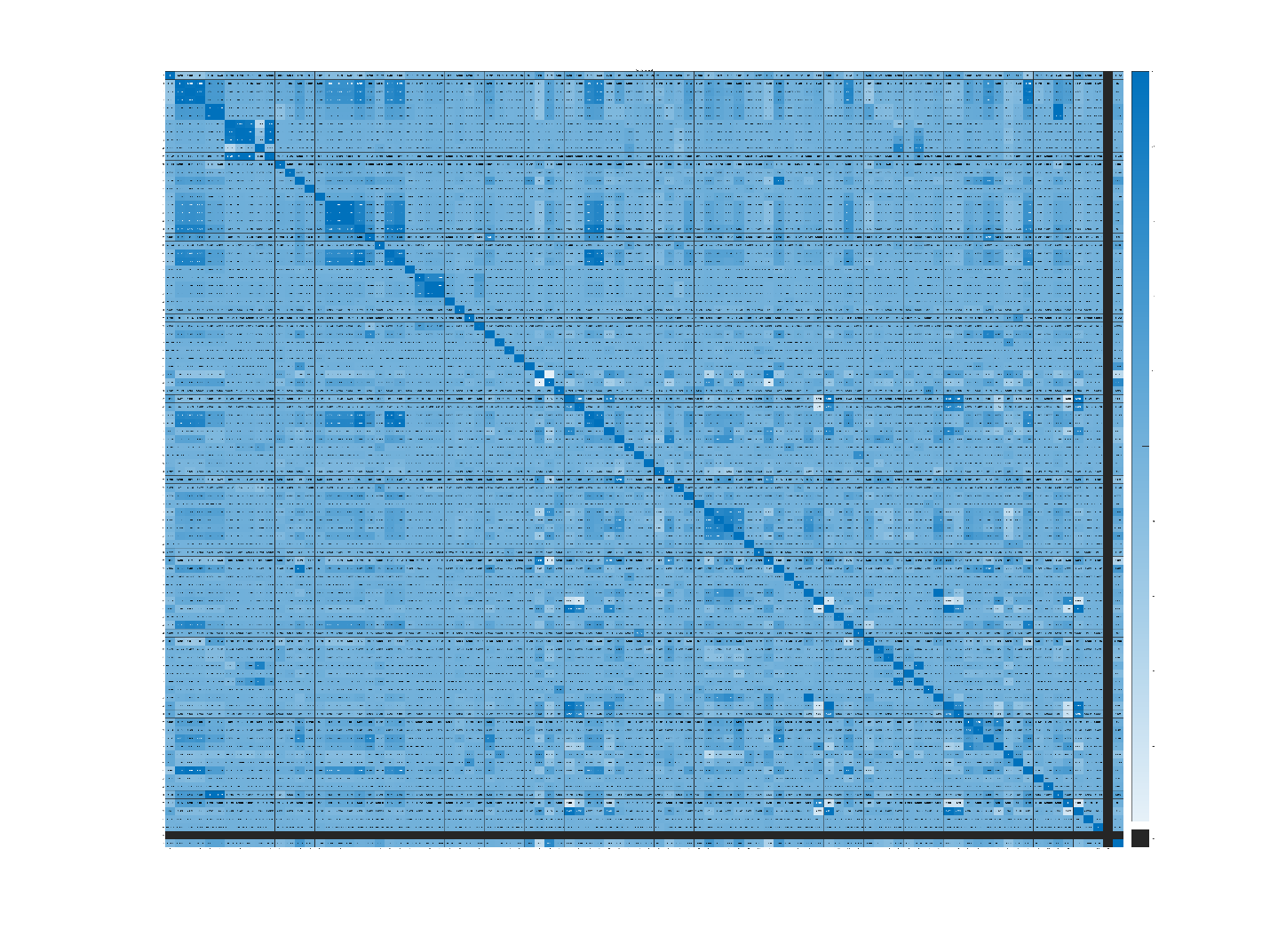
**2. Methodologies**

**2.1. Split the dataset**

We split the whole dataset into training and test dataset. 70% of data was used for training and 30% was used for test. We checked if there’s missing data in the dataset. As a result, there was no missing value in the whole dataset.

**2.2. Covariance analysis**

Correlation analysis is a statistical technique used to measure the strength and direction of the linear relationship between two or more variables. It is commonly used to assess the association between two continuous variables. The correlation plot is as follows.



As the figure shows, 95th variable does not have any correlation with the dependent variable.

**2.3. Logistic regression for all the predictors**

We trained the logistic regression model for the whole predictors. Logistic regression is a statistical method used for binary classification problems, where the dependent variable (target) is binary, meaning it can take only two possible values, such as 0 or 1. It is widely used in machine learning and statistics for tasks like predicting the probability of an event occurring.

Logistic regression is a simple and interpretable ML algorithm. It is efficient with small datasets and it’s robust to noise. But it has some cons such as it is limited to binary classifications. It needs the linearity assumption in the beginning so if the variables are non-linearly correlated, the result will be affected from this. It also has some drawbacks for imbalanced class distribution. This dataset is highly imbalanced so the result is affected by that.

In this step, we chose all the predictors and trained the logistic regression model for them.

**2.4. Logistic regression for most correlated predictors**

In this step, we chose 5 most highly correlated predictors and trained the logistic regression for them.

**2.5. Boosted classification tree**

A boosted classification tree, also known as a boosted tree or gradient boosting, is an ensemble learning method that combines multiple decision trees to create a powerful predictive model for classification tasks. It is a popular machine learning technique due to its high predictive accuracy and ability to handle complex relationships in the data.

It provides high accuracy and is robust to overfitting. And it can also handle non-linearity.

Disadvantage of this model is that it’s computationally intensive and as it has various hyperparameters, it takes so long for the optimization.

**2.6. Random Forest**

Random Forest is another popular ensemble learning method used for both classification and regression tasks. It is an extension of decision trees and addresses some of the limitations of a single decision tree by combining multiple trees to create a more robust and accurate model.

Random forest has several advantages such as high predictive accuracy, robustness, parallelizable and handling high-dimensional data. It also has some disadvantages such as complexity and memory usage.

**3. Main Findings**

We trained 2 logistic regression models and 1 boosted classification tree and 1 random forest for the dataset provided. For the boosted classification tree and random forest, hyper-parameter tuning is done using Bayesian optimization technique.

We calculated 2 main metrics from the models including accuracy and confusion matrix.

The result is as following:

- Logistic regression for all predictors

Accuracy Rate: 96.73%

Confusion matrix:   
 1979 0  
 67 0

- Logistic regression for top 5 predictors

Accuracy Rate: 96.48%

Confusion matrix:  
 1973 6  
 66 1

- Boosted classification tree

Accuracy Rate: 96.92%

Confusion matrix:  
 1974 5  
 58 9

- Random Forest

Accuracy Rate: 96.73%

Confusion matrix:  
 1969 10  
 57 10

The result shows that Logistic regression is not good enough for this problem. That’s because logistic regression has some drawbacks for imbalanced dataset. Top 5 predictors show the slightly better result because none values are excluded and only important variables are used for the training.

Boosted classification tree and Random Forest models show relatively good result for this imbalanced dataset.

**4. Conclusion**

In this project, we analyzed the bankruptcy of companies in the Taiwan Economic Journal. We ran a preliminary correlation analysis to get the visual findings of the dataset. After this, we split the dataset into training and test dataset and trained multiple models including Logistic regression, Boosted classification tree and Random forest models.

We fine-tuned the hyper parameters to get the best result for those models. After running the analysis with these models, we concluded that Boosted classification tree and random forest models show the better result for imbalanced data classification than logistic regression models.

We can utilize various techniques such as data augmentation to handle the class imbalance problem and get the better result.