Using Financial Data to Predict Bankruptcy

Z0173806

Data Source: https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction

Reliable bankruptcy prediction models open the door for governments and banks to predict future states of economies, improving policy formation/implementation. Machine learning techniques offer an alternative to traditional statistical methods of prediction such as discriminant analysis proposed by Altman (1968). This paper documents the implementation of a Logistic Regression binary classifier and proposes a Balanced Boosted Aggregating Logistic Regression classifier to create bankruptcy prediction models.

1. Introduction

1.1. Outputs and inputs

An 80% selection of the total dataset—hereafter "training set"—will be the input used by the models. The binary classifiers that are produced by the models are programmatically expressed as 0 or 1, with 0 being the negative class (solvency) and 1 the positive class (bankruptcy).

1.2. Motivation

Modelling insolvency is an essential tool for governments looking to assess economic activity, financial institutions minimising credit risks and small-scale lenders hoping to stimulate economic growth. Historical events such as the 2007/8 financial crisis indicate the monumental effect poor credit risk management has on the global economy.

2. Theoretical Background

2.1. Logistic Regression

For predicting bankruptcy, the aim of a logistic regression model is to use financial data features to construct a general hypothesis function $h(X_i)$, mapping the linear combination $(\Theta^T X_i)$, to the probability of a firm going bankrupt. The hypothesis function is trained by optimising the coefficient vector (Θ), such that the differences between the predicted probabilities $(\hat{y}_i)$, and actual classes (y_i) are minimised for all feature vectors (X_i) in the training set. Logistic Regression then uses a decision boundary to separate positive and negative classes.

$$h(X) = \frac{1}{1 + e^{-\Theta^T X}} \quad (1)$$

2.2. Stratified Cross Validation

Stratified Cross Validation (SCV) is an evaluation method used to assess how accurately a model predicts unseen data. In SCV, the dataset is divided into k subsets (folds) such that each fold has the same distribution of either class as the whole dataset. Folds are iteratively selected to be used as the validation set for the model trained on the other folds. This is repeated until all the folds have been used to validate a model. The accuracy data from this process is averaged to find the model's general accuracy.

2.3. Balanced Bootstrap Aggregating

Balanced BAgging (BB) is an ensemble method that combines under-sampling techniques with Bootstrap Aggregating (BAgging) to overcome bias associated with the imbalanced dataset problem. BB creates sample sets of the original data such that the share of the positive and negative classes is balanced (Gnip and Drotar, 2019). Starting with the initial training set, weights are applied, and a base learner hypothesis function is trained $(h_1(X))$. Using the results from $h_{i-1}(X)$ weights of misclassified datapoints are increased and $h_i(X)$ is trained $\forall\ i\ \in\ k$. The resultant data is aggregated to form a generalised hypothesis function $H(X)$.

$$H(X) = g\big(h_1(X), \quad h_2(X), \ldots, h_k(X)\big) \quad (2)$$

The aim of this technique is to use many weak classifiers to sequentially rectify the misclassifications made by the previous. This property suggests that BB, using Logistic Regression weak classifiers, should outperform SCV Logistic Regression. However, it is important to note that BB can be prone to overfitting, when too many weak classifiers are used.

3. Research Design

3.1. Data and Pre-processing

The data was collected from the Taiwan Economic Journal between 1999 and 2009. The features are ratios from the financial reports of the 6819 businesses. Since all company names have been

redacted, there is no way of knowing if the datapoints are all different companies. Some of the data may have been collected from the same firms at different times causing minimal variance between such instances, thus limiting the data's descriptive power.

As Logistic Regression models assume normally distributed data, the dataset was mean scaled (Hackeling, 2017).

The data was imbalanced in favour of the negative class (solvency) (figure 1). Applying a model to imbalanced data can result in overfitting when the underrepresented class does not appear in the training set. To ensure both the testing and training sets had a proportion of positive cases representative of the whole dataset, a stratified splitting method with no shuffling was used. 20% of the data was used for testing. Both models were trained and tested with common training and testing sets.

Information	Count
Number of instances	6819
Number of features	95
Positive Target (Bankrupt)	220
Negative Target (Not Bankrupt)	6599
Nan entries	0
Continuous Numerical Features	95

Figure 1: Descriptive statistics of the dataset.

Many of the features in the dataset are calculated with common variables, for example debt. This meant that a significant number of the variables were highly correlated with each other.

Analysis of Variance (ANOVA) feature selection was employed to eliminate features with low explanatory power. In the context of feature selection, ANOVA evaluates the explanatory power of the features by testing whether their means differ significantly. Features with low explanatory power are then selectively eliminated. Figure 2 shows the 10 features with the most explanatory power selected by the ANOVA selector. It can be estimated that decreasing the number of features will reduce the effects of overfitting by creating a simpler, more generalised model.

Features
ROA(C) before interest and depreciation before interest
ROA(A) before interest and % after tax
ROA(B) before interest and depreciation after tax
Persistent EPS in the Last Four Seasons
Per Share Net profit before tax (Yuan ¥)
Debt ratio %
Net worth/Assets
Net profit before tax/Paid-in capital
Retained Earnings to Total Assets
Net Income to Total Assets

Figure 2: ANOVA selected features. Features used in the model development process.

Figure 3 shows that there are many different correlations between the features. This indicates that the features have different correlations to the class.

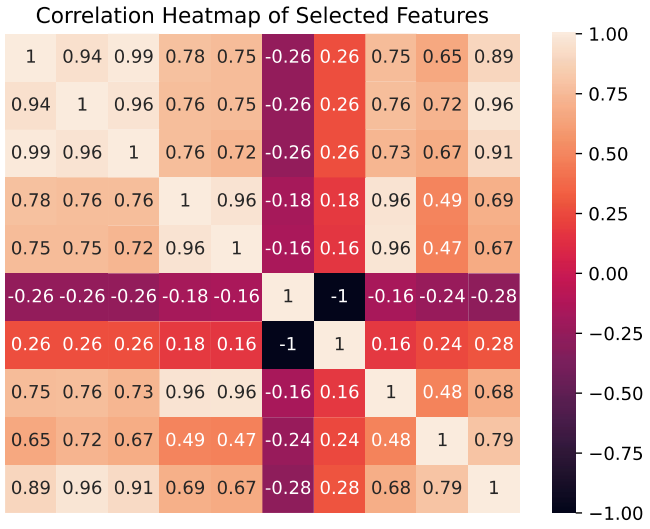

Figure 3: Pearson's Correlation heatmap. Rows from top to bottom are features as ordered in figure 2. Columns from left to right are features as ordered in figure 2.

4. Results

Appropriate performance metrics should be chosen depending on the intended application of the bankruptcy model. As the model was applied to a situation where both accurately predicting bankruptcy and predicting solvency are paramount, the performance metric used had to take into account precision and recall with equal importance. Due to the imbalanced nature of the dataset, F1-score was used to assess performance. Accuracy was deemed unacceptable because it is susceptible to influence from the more frequent class. Receiver Operating Characteristic (ROC) Curve analysis was also used in model comparison.

4.1. Logistic Regression Implementation

To tune SciKit-Learn's Logistic regression hyperparameters, a five-fold cross validated grid search was conducted to find the optimal combination of 'C' (the inverse regularisation strength, 'penalty' the Regularisation penalty and 'solver' the algorithm used in the model optimisation process). Hyperparameter combinations were ranked by f1-score.

	C	Penalty	Solver
Value	10	L2	liblinear

Figure 4: Combination of hyperparameters ranked highest by f1 score from a five-fold cross validated grid search.

Using the optimised parameters from the grid search the initial model was constructed. This model did not perform particularly well. To rectify this, in the second iteration of the model class, weighting was implemented by the method below such that the total number of incorrectly predicted positive classes incur the same cost as incorrectly predicted negative classes.

$$w_{positive} = \frac{n\_samples}{2 \times n\_positive} \quad (3)$$

$$w_{negative} = \frac{n\_samples}{2 \times n\_negative} \quad (4)$$

Figure 5 shows the benefits and drawbacks of unbalanced and balanced SCV logistic regression. Using balanced SCV regression does not necessarily make the selector any better at discerning between classes, it just encourages classifying more of the positive class. This explains the trade-off shown in figure 5 between precision and recall. Overall, the balanced model performed better than the unbalanced.

Score	Unbalanced SCV Logistic Regression	Balanced SCV Logistic Regression
F1 score	0.200000	0.272727
AUC	0.907283	0.913895
Precision	0.375000	0.163636
Recall	0.136364	0.818182

Figure 5: Model evaluation for class-weight balanced/ unbalanced five-fold stratified cross validated Logistic Regression

4.2. Balanced BAgging

Imbalanced Learn's Balanced BAgging was implemented with the optimised Logistic Regression model with no cross validation as the base learner. The number of samples that ought to be used in the final model was decided by conducting a search for the best performing model using one to ten samples. The maximum f1 score achieved in this search is shown in figure 6. For more than six samples the classifiers tend to worsen, due to overfitting. Six Base learners were used in the final model.

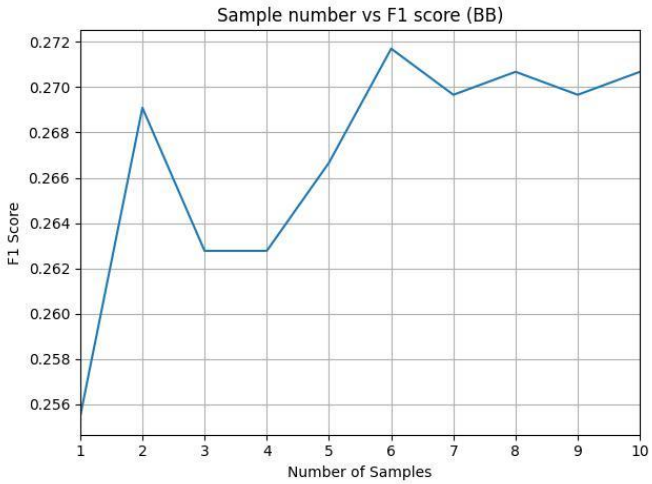

Figure 6: Sample number optimisation graph for Balanced BAgged Logistic Regression.

4.3. Comparison of Results

In both models precision is low. This suggests that positive predictions are not very reliable. However, recall is high, indicating that both models are effective at capturing the positive class. The AUC values are high indicating good discrimination between classes.

Score	5-Fold SCV Logistic Regression	Balanced BAgged
F1 score	0.272727	0.271698
AUC	0.913895	0.915152
Precision	0.163636	0.162896
Recall	0.818182	0.818182

Figure 7: Model evaluation for balanced five-fold stratified cross validated Logistic Regression and Balanced BAgged Logistic Regression.

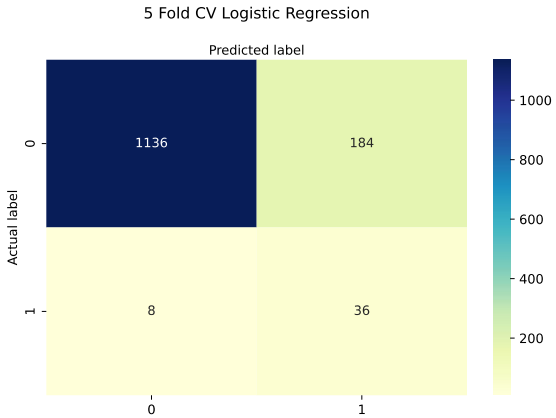

Figure 8: Confusion matrix for five-fold stratified cross validated Logistic Regression.

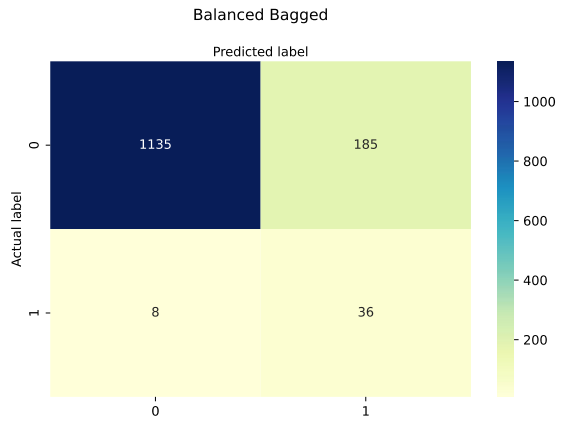

Figure 9: Confusion matrix for Balanced BAgged Logistic Regression.

Both models discriminate classes significantly better than random classification, evidenced by figure 10. The ROC curves of the two classifiers undulate out of phase in the middle part. This suggests that there is only a marginal difference at a few thresholds where one classifier can achieve a higher true positive rate for the same false negative rate.

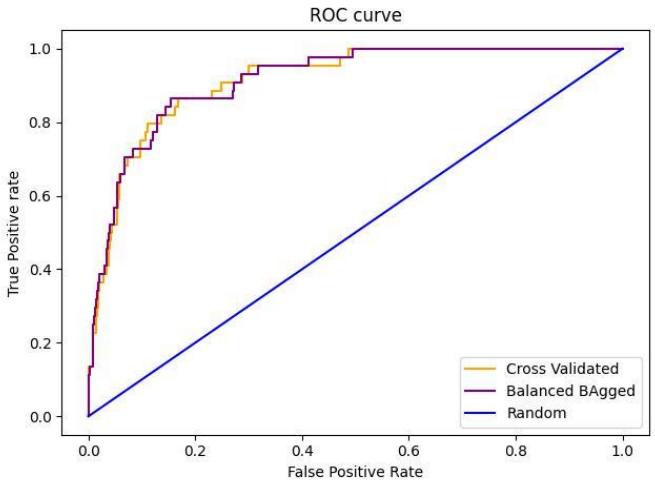

Figure 10: Receiver Operating Characteristic curves for five-fold stratified cross validated Logistic Regression and Balanced BAgged Logistic Regression.

5. Conclusion

In conclusion, the analysis of these models elucidates the challenges in predicting bankruptcy where successful businesses outnumber unsuccessful businesses in an economy. It is this disparity that causes difficulty in creating detailed models for predicting bankruptcy that accurately encompass the intricacies of bankruptcy indicators.

5-fold Stratified Cross Validated Logistic Regression performed marginally better than Balanced Bootstrap Aggregating using Logistic Regression, as the base learner.

It is likely that lack of depth of detail in the description provided by the models is due to utilising a narrow scope of data. The data used in this research did not contain any information about the age of the firms concerned or the economic climate at the time of reporting. Accommodating for these factors into the data would develop a more descriptive bankruptcy classification model. It would be more appropriate to construct sector specific models, as training a classifier on more homogeneous instances could increase the intricacy of the classifier's trend recognition.

6. Self-Evaluation

Throughout the Machine Learning Sub-module, I gained insights into the significance of simplicity in crafting quantitative models. In this assignment, I learnt to iteratively refine models for enhanced descriptiveness. While navigating the challenges, I found it challenging to strike a balance in evaluating models comprehensively without overwhelming with excessive information. Looking ahead, I would adopt a strategy of experimenting with many models before selecting one to optimise.

Reference list

Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, [online] 23(4), pp.589–609. Available at: https://www.jstor.org/stable/2978933 [Accessed 17 Jan. 2024].

Gnip, P. and Drotar, P. (2019). *Ensemble methods for strongly imbalanced data: bankruptcy prediction | IEEE Conference Publication | IEEE Xplore*. [online] ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9111557&tag=1 [Accessed 17 Jan. 2024].

Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn - Second Edition*. Second ed. Packt Publishing.