Comparison of Logistic Regression and Random Forest Classifier for Bankruptcy prediction

Description and motivation of the problem:

- The objective involves developing predictive models for a binary classification task targeting the likelihood of a company going bankrupt.
- The problem aims to use machine learning techniques, particularly logistic regression and random forest classifier, to forecast financial distress by using Taiwan's Economic data from the years 1999 to 2009 .
- The primary motivation is to create robust models that can accurately classify companies into bankrupt and non-bankrupt categories, aiding in risk assessment and proactive financial planning.

Initial analysis of the data set including basic statistics:

- The dataset for Taiwanese Bankruptcy Prediction is from UCI Machine Learning Repository.
- The dataset encompasses various financial features and indicators that are hypothesized to influence bankruptcy predictions.
- Dataset comprises 6819 rows and 96 columns.
- Predominantly numerical variables represent financial indicators and performance ratios.
- Various statistical summaries available for each numerical variable.
- Target variable named "Bankrupt" with values 0 and 1, indicating a possible binary classification.
- Subset of 20 features selected,(top 10 positively correlated and top 10 negatively correlated) including debt ratio, liability-to-assets ratios, profit margins, and return on assets.
- Class imbalance noted before SMOTE: 6599 instances of Class 0, 220 instances of Class 1. After SMOTE, both classes balanced to 6599 instances each.
- No missing values detected across features or columns.
- No duplicate rows found in the dataset.

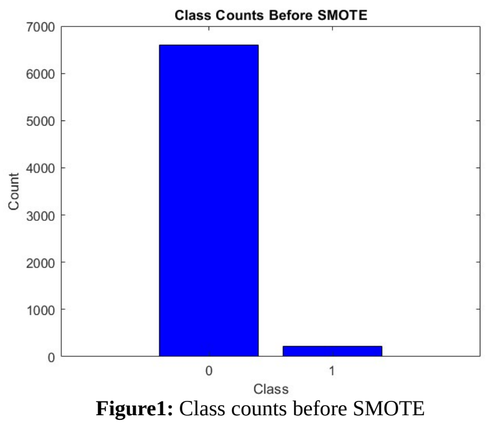

Figure1: Class counts before SMOTE

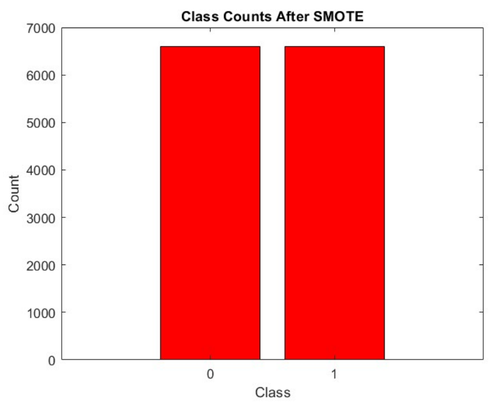

Figure 2: Class counts after SMOTE

Column_Name	Min	Max	Standard_Deviation	Mean
{'NetIncomeFlag'}	0	1	0.077543	0.46334
{'DebtRatio_'}	0	1	0.092544	0.51057
{'CurrentLiabilityToAssets'}	0	1	0.083154	0.50931
{'CurrentLiabilityToCurrentAssets'}	0	1	0.030726	0.17618
{'TotalExpense_Assets'}	0	1	0.030814	0.17612
{'BorrowingDependency'}	0	1	0.036203	0.20938
{'FixedAssetsTurnoverFrequency'}	0	1	0.034225	0.16673
{'LiabilityToEquity'}	0	1	0.063005	0.14853
{'CurrentLiabilities_Equity'}	0	1	0.063005	0.85147
{'CurrentLiabilityToEquity'}	0	1	0.044406	0.38253
{'NetWorth_Assets'}	0	1	0.032423	0.16579
{'ROA_C_BeforeInterestAndDepreciationBeforeInterest'}	0	9.99e+09	2.8147e+09	1.4593e+09
{'PersistentEPSInTheLastFourSeasons'}	0	1	0.057125	0.11674
{'ROA_B_BeforeInterestAndDepreciationAfterTax'}	0	1	0.039576	0.33719
{'NetProfitBeforeTax_Paid_inCapital'}	0	1	0.037345	0.038604
{'ROA_A_BeforeInterestAnd_AfterTax'}	0	1	0.039576	0.33719
{'PerShareNetProfitBeforeTax_Yuan__'}	0	1	0.035486	0.044327
{'NetIncomeToTotalAssets'}	0	1	0.064899	0.77494
{'NetValuePerShare_B_'}	0	1	0.041281	0.28693
{'NetValuePerShare_A_'}	1	1	0	1

Table1: Description of selected features after normalization

The summary of implementing model with their pros and cons:

Logistic Regression:

Logistic Regression is a go-to for simpler analyses, emphasizing interpretability, and when a linear relationship between features and outcome is assumed.

Pros:

- Simplicity: Easy to implement and interpret.
- Provides probabilities for outcomes.
- Works well with smaller datasets.

Cons:

- Assumes linear relationship between features and outcome.
- May not capture complex patterns in data well.
- Sensitive to outliers and multicollinearity.

Random Forest:

Random Forest excels in handling complex, non-linear data relationships and achieving higher accuracy, but might be less interpretable due to its ensemble nature and computational cost.

Pros:

- Handles non-linear relationships in data effectively.
- Robust to overfitting.
- Works well with large datasets and high-dimensional spaces.
- Provides feature importance.

Cons:

- Can be computationally expensive.
- May not be as easily interpretable as simpler models.
- Black-box nature: Understanding the model's decision-making process can be challenging.

Hypothesis statement:

- logistic regression and random forest classifiers will demonstrate varying predictive performances in determining the likelihood of a company facing financial distress.
- random forest classifier will outperform logistic regression in accurately categorizing companies as bankrupt or non-bankrupt based on the financial indicators provided in the dataset.
- This expectation is grounded in the assumption that the random forest model's ability to handle complex relationships and interactions among features will result in superior predictive power compared to logistic regression, ultimately leading to higher accuracy and better overall performance metrics.
- Implementation of logistic regression and random forest classifiers on the dataset featuring various financial indicators and company attributes will demonstrate that machine learning models can effectively predict the likelihood of a company facing bankruptcy.

Description of the choice of training and evaluation methodology:

- Prepare data by segregating features and target variables, then visualize initial class distribution to comprehend inherent imbalances.
- Utilize SMOTE to rectify class imbalance, employing K-nearest neighbors (KNN) to generate synthetic samples for the minority class and balance class representation.
- Conduct exploratory data analysis to determine feature-target correlations, identifying top positively and negatively correlated features.
- Select the top 10 positively and negatively correlated features, refining the dataset for further analysis.
- Train Logistic Regression and Random Forest models using 5-fold cross-validation for model evaluation.
- Evaluate model performance with metrics like Accuracy, Precision, Recall, and F1 Score across different data subsets (folds).
- Choose the best models based on their F1 Score across folds for further consideration.
- Generate training and testing error curves for Logistic Regression and Random Forest models to visualize their performance.
- Calculate and present consolidated average performance metrics obtained from the folds for a comprehensive view of model effectiveness.
- Visualize confusion matrices to assess the overall predictive capabilities of the models across the entire dataset.
- Save the best Logistic Regression and Random Forest models for future deployment or analysis.

Choice of parameters and experimental results:

SMOTE Parameters:

`num_synthetic_samples`: Calculated as the difference between the number of majority and minority samples (6379).
`knn_model = fitcknn(X(y == 1, :), y(y == 1), 'NumNeighbors', 5)`: `NumNeighbors` is set to 5 for the K-nearest neighbors algorithm.

Logistic Regression Model Parameters:

`Lambda`: Regularization parameter set to 0.01.
`Standardize`: True for scaling the predictors.
`Alpha`: Set to 1, indicating LASSO regularization.
`fitglm`: Logistic regression model fitting with specified parameters: `'Distribution' = 'binomial'`, `'Link' = 'logit'`.

Random Forest Model Parameters:

`numTrees = 100`: Number of trees in the Random Forest classifier.

LR	Model	RF
0.88556	Accuracy	0.95682
0.8703	Precision	0.94126
0.9036	Recall	0.97722
0.88664	F1-Score	0.9589

Table2: Metrics for both models

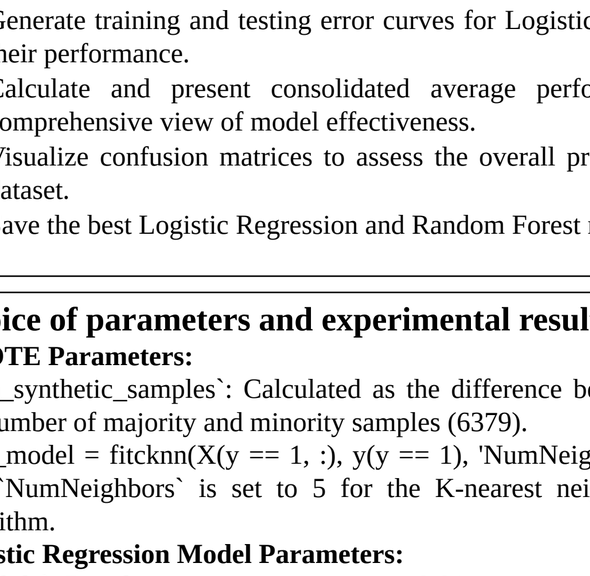

Figure6: Performance across folds for both models

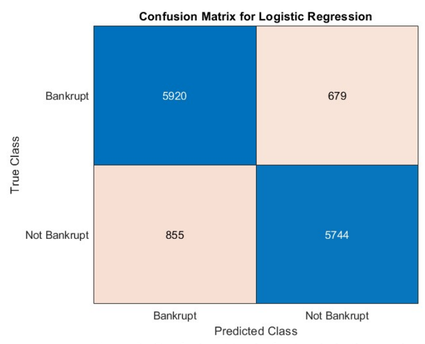

Figure3: Confusion Matrix for Logistic Regression

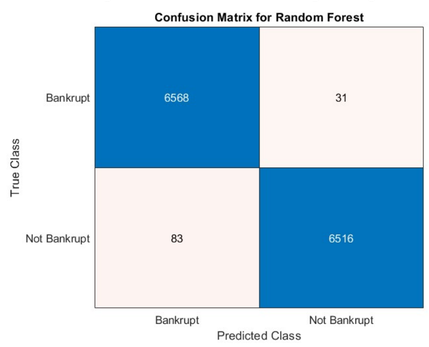

Figure4: Confusion Matrix for Random Forest

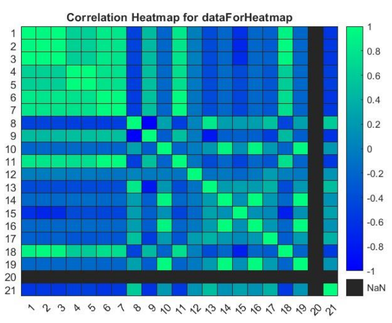

Figure5: Corelation Heatmap for selected features

Analysis and critical evaluation of results:

- Both models, logistic regression and random forest, performed well across all folds. Random forest slightly outperformed, averaging 95.68% accuracy, while logistic regression achieved 88.55%. This indicates their ability to learn data relationships and make accurate predictions.
- Logistic regression displayed good precision at 87.03%, commendable given its simplicity. Yet, Random Forest surpassed LR with 94.12% precision.
- Logistic regression had an average recall of 90.36%, slightly higher than precision. Meanwhile, random forest achieved a higher recall of 97.72%, surpassing its precision.
- F1 score for logistic regression was 88.66%, while random forest achieved a higher score of 97.5%, signifying a good balance between precision and recall for both models.
- Logistic regression correctly predicted 90.1% of bankrupt cases and 87.1% of non-bankrupt cases. However, it showed a tendency for false positives (10.3%) over false negatives (12.9%).
- Reducing false positives in logistic regression by using a more stringent threshold might increase false negatives, a trade-off to consider favoring fewer false negatives.
- Random forest exhibited high accuracy: 98.9% for bankrupt and 99.0% for non-bankrupt cases, with remarkably low false positive (0.5%) and false negative (0.3%) rates.
- Improvement for the random forest lies in interpretability, a known challenge for this model. Techniques like partial dependence plots and permutation importance could enhance interpretability.
- This condensed version emphasizes the comparative performance of both models, their predictive accuracy, and the interpretability challenge specifically for the random forest model.

Lessons learned, future work and references:

Lessons learned:

- The trade-off between model interpretability and performance should be considered based on the application's specific needs and constraints.
- Understanding the dataset's quality and ensuring it represents diverse samples is critical to avoid bias and ensure model generalizability.
- It's crucial to consider computational efficiency, interpretability, and scalability when choosing between models, especially if deployed in a real-world scenario.
- Balancing techniques like SMOTE are crucial to prevent model bias towards the majority class.
- Properly handling column headers and data types is essential for smooth analysis.

Future Work:

- Robustness testing on unseen or future data is essential to ensure the model's reliability and generalizability beyond the current dataset.
- Investigate and resolve problems related to multicollinearity or rank deficiency in the regression design matrix to improve model stability.
- Optimize code and models for scalability, especially if dealing with larger datasets, to ensure computational efficiency.
- Translating model performance into actionable business decisions is vital. Understanding the impact of false positives and false negatives on financial decisions is crucial.
- Make use of predictor importance estimation to select features, while training Random Forest to check if that model performs better than the model trained with current features.

References:

- Taiwanese Bankruptcy Prediction. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D.
- Hauser, R.P. and Booth, D., 2011. Predicting bankruptcy with robust logistic regression. Journal of Data Science, 9(4), pp.565-584.
- Chih-Fong Tsai,Feature selection in bankruptcy prediction,Knowledge-Based Systems,Volume 22, Issue 2,2009,Pages 120-127,ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2008.08.002
- Yi Qu, Pei Quan, Minglong Lei, Yong Shi,Review of bankruptcy prediction using machine learning and deep learning techniques,Procedia Computer Science,Volume 162,2019,Pages 895-899,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.12.065 .
- S. Joshi, R. Ramesh and S. Tahsildar, "A Bankruptcy Prediction Model Using Random Forest," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 1-6, doi: 10.1109/ICCONS.2018.8663128.
- Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, Guan-An Shih,Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study,European Journal of Operational Research,Volume 252, Issue 2,2016,Pages 561-572,ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2016.01.012 .