Company Bankruptcy Prediction

Management / Research Question:

Can a company's financial ratios be used to predict whether it will go bankrupt in the near future? This question matters because early detection of financial distress helps managers, investors, and lenders make better decisions. If a firm's likelihood of bankruptcy can be predicted accurately, executives can take corrective actions such as restructuring debt, adjusting strategy, or cutting costs. Likewise, banks and investors can use the predictions to manage lending risk and protect capital, while policymakers can use them to monitor systemic stability in the broader economy

Data Preparation, Exploration, and Visualization:

The dataset used for this analysis comes from Kaggle's Taiwan Company Bankruptcy Prediction repository and contains 6,819 companies observed between 1999 and 2009, each described by 95 financial ratios. The target variable, Bankrupt?, is binary (1 = bankrupt, 0 = not bankrupt), and only 3.2% of companies were labeled bankrupt, creating a substantial class imbalance. No missing values were present, and all predictors were numeric, allowing for straightforward processing. All features were standardized using z-scores to equalize scale across variables, and any zero-variance columns were dropped to improve model stability. Exploratory analysis showed that bankrupt firms generally had lower profitability, higher leverage, and weaker liquidity compared to solvent firms, trends that align with known financial-distress indicators.

Research Design and Modeling Methods:

The objective of this study was to predict company bankruptcy using financial ratios as predictors and to compare the performance of three ensemble learning models. The dataset was split using stratified sampling to preserve the ratio of bankrupt to non-bankrupt firms, with 80% of the data used for training and 20% for validation. Each model was implemented within a scikit-learn pipeline that handled numeric preprocessing, including median imputation (where needed) and standardization. Three ensemble tree classifiers were trained and tuned: Random Forest, Gradient Boosting, and Extra Trees. Each model underwent 5-fold cross-validation, and hyperparameters were optimized for n_estimators, max_depth, max_features, and criterion (entropy or gini). Because the data were highly imbalanced, models were trained using class_weight='balanced', and the F1-score was chosen as the main performance metric since it captures the trade-off between precision and recall.

Results and Model Evaluation:

Model performance was assessed using the F1-score on a 20% validation set, which balances precision (avoiding false positives) and recall (capturing true bankruptcies). Each model's hyperparameters were tuned through 5-fold cross-validation, and their performance was compared on both the training and validation sets. The Extra Trees Classifier achieved the best overall results, with a cross-validated F1-score of 0.407 and a validation F1-score of 0.485, correctly identifying approximately 75% of bankrupt firms (recall) while maintaining a precision of 0.36. The Random Forest model performed slightly worse (validation F1 = 0.452), while Gradient Boosting achieved similar results (validation F1 = 0.468). All models produced high overall accuracy (~96%), but accuracy was misleading due to the class imbalance. F1-score provided a more realistic view of predictive performance, confirming that Extra Trees offered the best trade-off between recall and precision.

Overall Results Summary:

All three ensemble models Random Forest, Gradient Boosting, and Extra Trees significantly outperformed a random baseline and demonstrated strong generalization between cross-validation and validation scores. The Extra Trees Classifier provided the best balance between sensitivity and precision, confirming its value as an early warning model for financial distress. It predicted 92 firms as bankrupt, compared to 44 actual bankruptcies in the validation data, capturing most of the true bankruptcies but generating some false alarms. This higher recall is desirable for risk management applications, where missing a failing firm is far more costly than investigating a false positive. Across all models, the most influential features included profitability and leverage ratios such as Net Income to Total Assets, Debt to Equity, and Operating Margin, aligning closely with established financial theory on corporate distress.

Output:

Training RandomForest...
Fitting 5 folds for each of 24 candidates, totalling 120 fits

Best params for RandomForest: {'model__criterion': 'entropy', 'model__max_depth': 10, 'model__max_features': 0.5, 'model__n_estimators': 500}
Best CV F1-score: 0.4170
Validation F1-score: 0.4516

	precision	recall	f1-score	support
0	0.9825	0.9788	0.9806	1320
1	0.4286	0.4773	0.4516	44
accuracy			0.9626	1364
macro avg	0.7055	0.7280	0.7161	1364
weighted avg	0.9646	0.9626	0.9636	1364

Training GradientBoosting...
Fitting 5 folds for each of 8 candidates, totalling 40 fits

Best params for GradientBoosting: {'model__max_depth': 2, 'model__max_features': 'sqrt', 'model__n_estimators': 500}
Best CV F1-score: 0.3745
Validation F1-score: 0.4675

	precision	recall	f1-score	support
0	0.9805	0.9886	0.9845	1320
1	0.5455	0.4091	0.4675	44
accuracy			0.9699	1364
macro avg	0.7630	0.6989	0.7260	1364
weighted avg	0.9664	0.9699	0.9679	1364

Training ExtraTrees...

```
Fitting 5 folds for each of 24 candidates, totalling 120 fits

Best params for ExtraTrees: {'model__criterion': 'entropy', 'model__max_depth': 10,
'model__max_features': 0.5, 'model__n_estimators': 600}
Best CV F1-score: 0.4072
Validation F1-score: 0.4853
              precision    recall  f1-score   support

           0     0.9914    0.9553    0.9730      1320
           1     0.3587    0.7500    0.4853        44

    accuracy                         0.9487      1364
   macro avg     0.6750    0.8527    0.7291      1364
weighted avg     0.9709    0.9487    0.9573      1364


=== Model Comparison ===
              Model  Best CV F1  Validation F1  \
2        ExtraTrees    0.407158       0.485294
1  GradientBoosting    0.374477       0.467532
0      RandomForest    0.417042       0.451613

                                          Best Params
2  {'model__criterion': 'entropy', 'model__max_de...
1  {'model__max_depth': 2, 'model__max_features':...
0  {'model__criterion': 'entropy', 'model__max_de...
```