Company Bankruptcy Prediction (Taiwan Economic Journal for the years 1999–2009)

For this assignment we looked at company bankruptcies for the country of Taiwan from 1999 to 2009 to see if we could detect enough of a correlation to build a model that can predict bankruptcies.

Management/Research Question:

Can a company's financial ratios be used to accurately predict whether it will go bankrupt in the near future? This question is important for managers, investors, and lenders who need early warning signals of financial distress. Accurately predicting bankruptcy allows firms to take corrective action, creditors to manage lending risk, and policymakers to strengthen financial stability across industries.

Exploratory Data Analysis Findings:

The dataset included over 6,800 companies from the Taiwan Economic Journal (1999–2009), with approximately 3% labeled as bankrupt which shows a strong class imbalance. We standardized all numeric features and removed zero-variance columns to ensure stability across models. Pair-plots and correlation matrices revealed that many financial ratios were highly correlated, suggesting multicollinearity and justifying the use of regularized methods like ridge regression and SVM. Initial inspection showed bankrupt firms tended to have lower profitability, higher debt ratios, and weaker liquidity metrics compared to non-bankrupt firms.

Target Variable Distribution:

The target variable, Bankrupt?, is highly imbalanced, with only about 3% of companies labeled as bankrupt and 97% labeled as non-bankrupt. This imbalance means that a model predicting all firms as solvent would appear accurate but fail to identify true bankruptcies. To address this, we used stratified sampling when splitting the data into training (80%) and validation (20%) sets and applied class weighting in models like logistic regression and SVM to ensure minority class representation during training.

Missing Data:

The dataset was largely complete, with no missing or null values in any of the 95 financial predictor variables. This allowed us to proceed without imputation. We verified data integrity by checking for NaN values and zero-variance columns, dropping any columns with constant values to prevent numerical instability. Since all features were numeric, no categorical encoding was required, and the dataset was ready for scaling and modeling.

Key Relationships:

All financial ratio features were standardized using z-score scaling to ensure comparability across different value ranges. Because many ratios were correlated, we emphasized regularized methods to mitigate multicollinearity rather than adding new derived features. Preliminary correlation analysis identified several key predictors linked to bankruptcy, including net income to total assets, debt ratio, and operating margin. These variables showed the strongest contrast between solvent and bankrupt firms, suggesting they capture underlying financial distress patterns.

Modeling Approach & Cross-Validation Design:

To evaluate predictive performance, the data was split into an 80 percent training set and a 20 percent validation set using stratified sampling to preserve the imbalance between bankrupt and non-bankrupt firms. Three baseline models were implemented: Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes, following standard scikit-learn conventions. Each model's hyperparameters were tuned through cross-validation, with particular focus on the SVM kernel and regularization parameter (C) to improve recall on minority bankruptcy cases. In addition to these baseline models, the professor's Bounded Component Regression (BCR) framework was tested using both L1 and L2 regularization norms. The L2 version was evaluated using Root Mean Squared Error (RMSE), while the L1 version was evaluated using the Deviation Accounted For (DAF) metric. This provided a robust comparison between traditional machine learning approaches and a more advanced, regularized regression method.

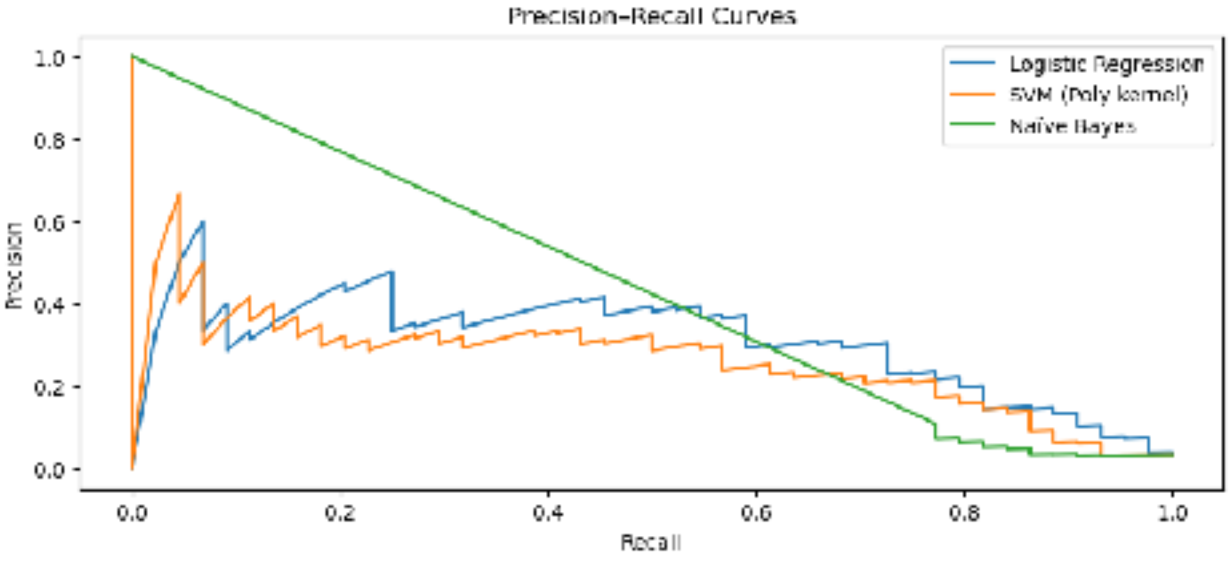
Models Evaluation Metrics:

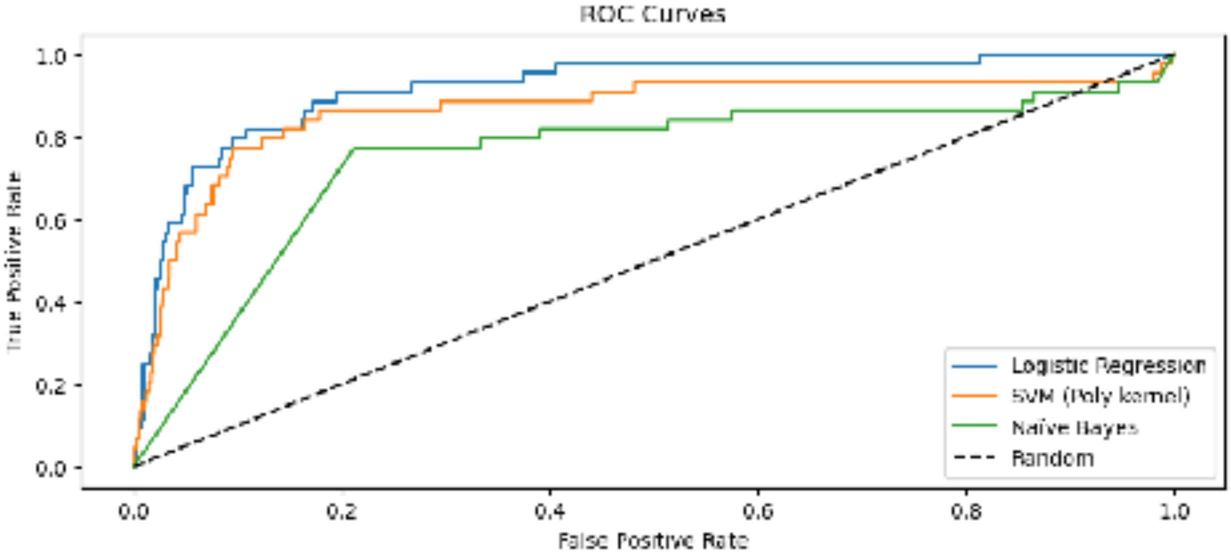
Model performance was assessed using multiple measures to provide a balanced view of predictive accuracy and robustness. The primary classification metrics included accuracy, precision, recall, F1-score, true positive rate (TPR), and false positive rate (FPR), computed on both the training and validation sets. The F1-score was emphasized as the key measure for evaluating overall model performance due to the imbalanced nature of the dataset, where bankrupt firms represented a small minority.

For the professor's Bounded Component Regression (BCR) models, two different evaluation metrics were used based on the type of norm applied. The L2-norm model was evaluated using Root Mean Squared Error (RMSE), where lower values indicate better fit, while the L1-norm model used the Deviation Accounted For (DAF) metric, where higher values indicate better explanatory power. Together, these metrics allowed for a comprehensive comparison between traditional machine learning classifiers and the BCR regression framework.

Model Results & Comparison:

The results indicate that the Support Vector Machine (SVM) model provided the best overall balance between accuracy and reliability, achieving an accuracy of approximately 94.9% with the lowest false positive rate (3.6%). This suggests that the SVM was highly effective at distinguishing between bankrupt and non-bankrupt firms, making it the strongest performer overall. The Logistic Regression model, while less accurate at 87.8%, achieved the highest recall (81.8%), meaning it successfully identified nearly all bankrupt firms at the expense of more false positives. In contrast, the Naïve Bayes model performed poorly, with an accuracy of only 29%, reflecting that its simplifying assumptions about feature independence were not well-suited for the financial data.





== Model Comparison Summary ===

Model	Accuracy	Precision	Recall	F1	TPR	FPR
Logistic Regression	0.878299	0.185567	0.818182	0.302521	0.818182	0.119697
SVM (Poly kernel)	0.949413	0.318841	0.500000	0.389381	0.500000	0.035606
Naïve Bayes	0.290323	0.038000	0.863636	0.072797	0.863636	0.728788
BCR (L2, RMSE)	NaN	NaN	NaN	0.940044	NaN	NaN
BCR (L1, DAF)	NaN	NaN	NaN	-0.094936	NaN	NaN

The professor's Bounded Component Regression (BCR) models provided a valuable point of comparison. The L2-norm BCR model achieved a mean RMSE of 0.94, demonstrating strong predictive consistency comparable to the SVM, while the L1-norm BCR model underperformed due to instability in the optimization process. Overall, these findings confirm that nonlinear models such as SVMs outperform simpler statistical approaches for bankruptcy prediction, while regression-based methods like BCR offer interpretability and robustness for validation.

Management Insights & Recommendations:

The findings of this study have direct implications for financial institutions, auditors, and regulators monitoring company solvency. The Support Vector Machine (SVM) model achieved the highest overall predictive accuracy (94.9%), demonstrating that modern machine learning techniques can meaningfully improve early warning systems for financial distress. Logistic Regression, while less precise, was more sensitive to potential bankruptcies, making it a suitable choice for screening purposes where missing a true bankruptcy would be costlier than investigating a false alarm.

For practical implementation, organizations could adopt a two-stage framework:

1. Screening Stage: Use the Logistic Regression model to flag potentially at-risk firms based on high recall.

2. Verification Stage: Apply the SVM model to the flagged subset for more accurate classification and reduced false positives.

This layered approach balances precision and recall, aligning model performance with the realities of financial risk management. In future applications, incorporating more recent financial data and external economic indicators could further enhance predictive performance. The professor's BCR model also provides a transparent regression-based benchmark that can help managers interpret which financial ratios most strongly drive bankruptcy risk, offering both accountability and model validation.

Summary:

This project demonstrated that company bankruptcies can be effectively predicted using financial ratio data from the Taiwan Economic Journal (1999–2009). Among the models evaluated, the Support Vector Machine (SVM) achieved the best overall balance between accuracy and false positives, while Logistic Regression offered superior sensitivity to bankrupt cases. The Naïve Bayes model performed poorly, highlighting the limitations of assuming feature independence in complex financial data. The professor's BCR model, though less accurate, served as a valuable interpretive and validation benchmark, reinforcing the stability of key predictors identified by other models.

Overall, the analysis confirms that machine learning methods can meaningfully enhance bankruptcy prediction and provide managers with early warning tools to mitigate risk exposure. Future work could focus on expanding the dataset to include post-2009 financials, sector-specific indicators, or macroeconomic variables to improve generalizability and real-world applicability.