COMPANY BANKRUPTCY PREDICTION

CREATED BY: [GROUP 21]
Zeel Patel (A20556822)
Ruchika Rajodiya (A20562246)
Dhruvi Pancholi (A20545574)
Sundar Machani (A20554747)

1. Introduction

Background:

Financial stability is crucial for business success, as insolvency can lead to investor losses, layoffs, and supply chain disruptions. Accurate bankruptcy predictions enable timely interventions and prevent such outcomes. Traditional methods often miss subtle financial patterns, relying on linear models or intuition. Machine learning addresses these limitations, offering improved accuracy and deeper insights. Advanced algorithms uncover hidden trends, enhance predictions, and build resilience against uncertainties.

Objective:

This project uses machine learning for bankruptcy prediction through:

- Testing multiple models (Logistic Regression, Random Forest, XGBoost, KNeighbourClassifier, Deep Learning)
- Developing accurate and balanced models
- Ensuring model reliability and practical application
- Providing clear, interpretable results for financial decision-making and strategy

2. Data

2.1 Dataset Description

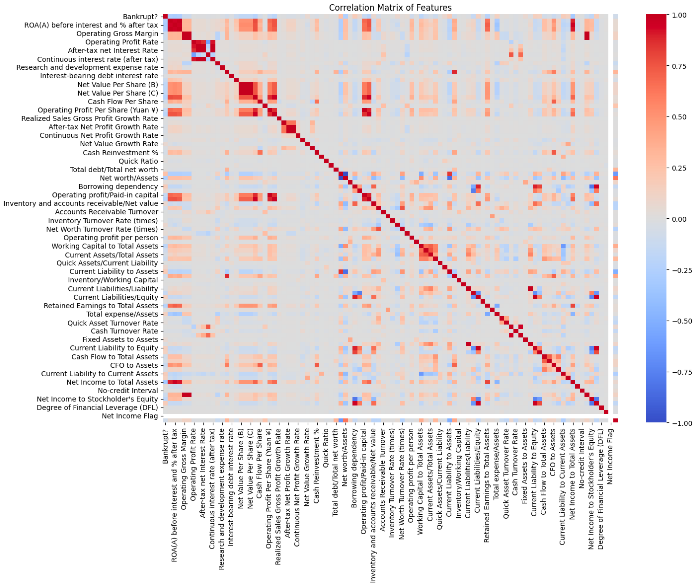
Dataset Overview:

The Company Bankruptcy Prediction dataset contains financial attributes with a binary target (‘Bankrupt?’), designed to classify companies as either financially stable or bankrupt. It includes 95 financial features (X1-X95), covering critical aspects of business performance such as return ratios, profit margins, liquidity metrics, leverage indicators, and efficiency ratios. These features provide a comprehensive view of a company's financial health, allowing for detailed analysis and accurate prediction. The dataset offers a rich source of information to evaluate the complex interplay between financial attributes and insolvency risk, making it ideal for machine learning applications.

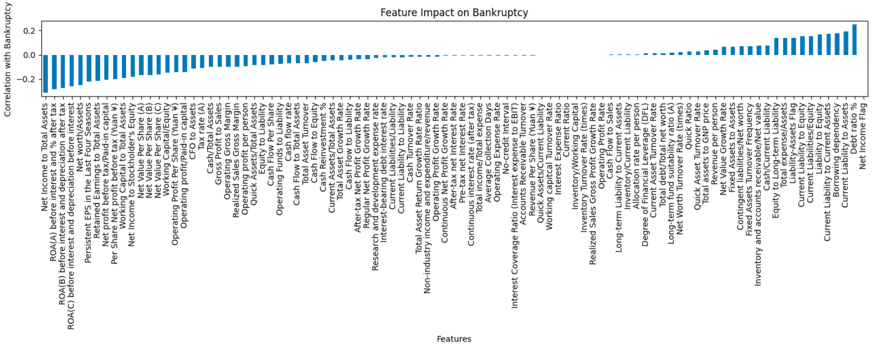
2.2 Data Processing:

- Label encoding for categorical variables
- StandardScaler for numerical features
- SMOTE for handling class imbalance (96.77% stable, 3.23% bankrupt)
- Correlation analysis visualized via a heatmap to explore feature relationships

- *Correlation Analysis:* Heatmap explores relationships between features.

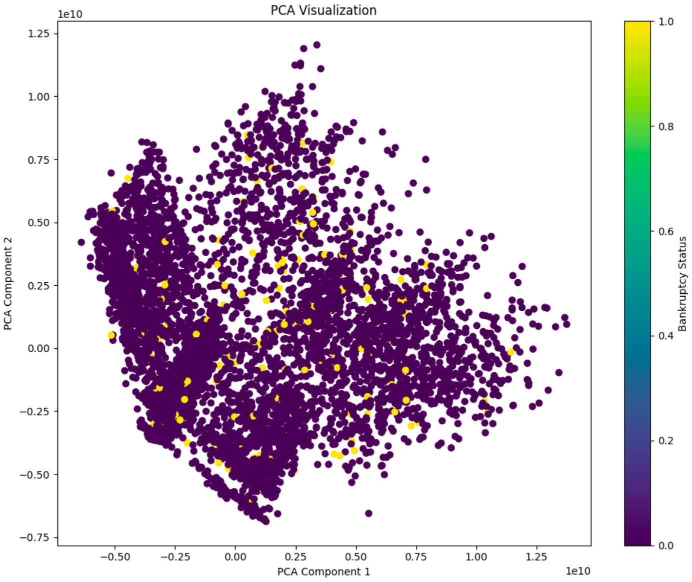


- *Feature Impact on Bankruptcy:* Bar chart visualizes the impact of features on bankruptcy.

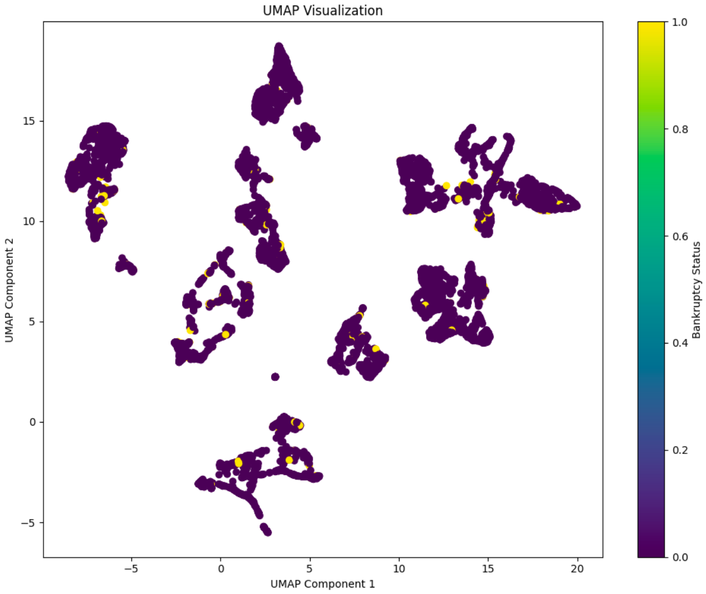


2.3 Dimensionality Reduction Techniques:

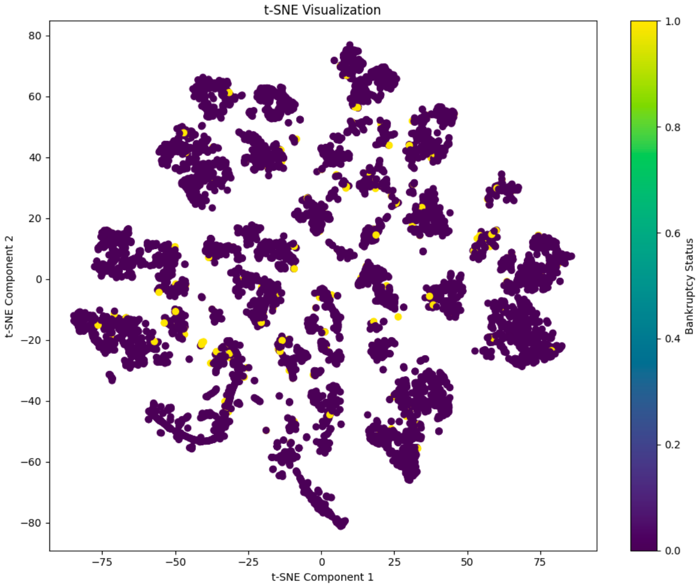
- *PCA Visualization:* Reduces dimensionality to two components.



- *UMAP Visualization:* Provides insights into the data structure.



- *t-SNE Visualization:* Captures local relationships in a 2D space.



Model	Validation Accuracy	Cross-Validation Mean Accuracy	Test Accuracy	Precision (0/1)
Logistic Regression	89%	89%	89%	0.90/0.88
Random Forest	97%	97%	97%	0.99/0.95
XGBoost	98%	98%	98%	1.00/0.96
KNeighbourClassifier	93%	93%	93%	1.00/0.88
Deep Learning	98%	97%	98%	1.00/0.96

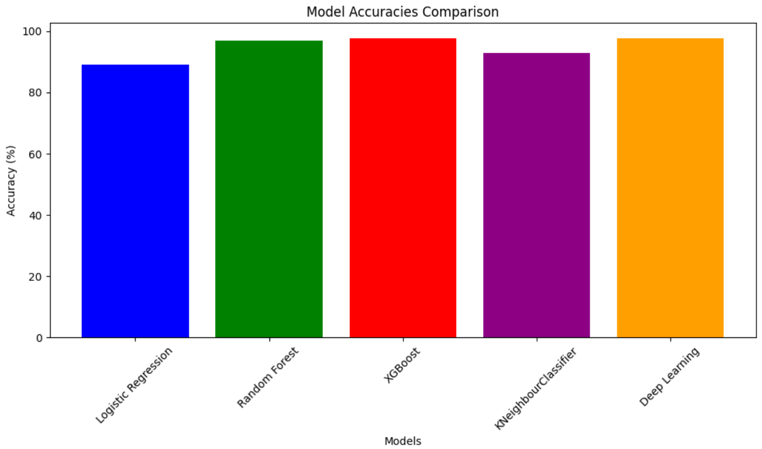
Key Findings:

- Correlation matrix and feature analysis provided initial insights
- Dimensionality reduction revealed data distribution patterns

- Visualizations guided model development

Model Performance:

XGBoost achieved the highest accuracy, followed closely by Random Forest and Deep Learning models.



Accuracy Plot of all models

4. Model Evaluation

4.1 Cross-Validation

Cross-validation is essential for evaluating model robustness. A 5-fold cross-validation approach is used to strike a balance between computational efficiency and reliable performance estimates. This method helps detect issues such as overfitting or underfitting by training the model on different data subsets.

4.2 Performance Comparison

The models—Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors (KNN), and Deep Learning (MLPClassifier)—are assessed based on key metrics: accuracy, precision, recall, and F1-score. The evaluation highlights each model's strengths and weaknesses, guiding the selection of the most appropriate model for the task.

5. Potential Improvements

5.1 Identified Pitfalls

Convergence warnings were observed in the Deep Learning model (MLPClassifier), which could impact its predictive performance.

5.2 Proposed Improvements

Deep Learning Model (MLPClassifier):

- **Increase Maximum Iterations:** Adjust the max_iter parameter to address convergence issues.
- **Hyperparameter Tuning:** Fine-tune hyperparameters to improve convergence and overall performance.

General Recommendations:

- **Ensemble Methods:** Explore advanced ensemble techniques like stacking or blending to boost accuracy.
- **Feature Engineering:** Experiment with creating or transforming features to enhance model performance.

5.3 Additional Experiments

Algorithm Variations:

- Test alternative algorithms or variations to gain a broader understanding of model behavior.

Feature Selection Methods:

- Investigate feature selection techniques beyond Logistic Regression to refine the input variables.

These experiments aim to improve predictive performance and provide deeper insights into both the dataset and model behavior.

6. Code and Dataset

Code Repository: https://github.com/sundarmachani-2752/CSP-571-DPA-Project/
Dataset: https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction/data

7. Conclusion

This project focused on applying machine learning techniques for bankruptcy prediction, offering valuable insights into companies' financial health.

Key Highlights:

1. **Dataset Exploration:** The 'Company Bankruptcy Prediction' dataset provided a rich set of financial attributes. Preprocessing steps included label encoding, feature scaling, and addressing class imbalance using SMOTE.
2. **Model Development:** Various algorithms were explored, including Logistic Regression, Random Forest, XGBoost, KNN, and Deep Learning (MLPClassifier). Logistic Regression also aided in feature selection.
3. **Model Evaluation:** A robust cross-validation strategy was employed to thoroughly evaluate models. Comparative analysis revealed each algorithm's strengths and weaknesses.

Final Thoughts:

This project establishes a strong foundation for using machine learning in bankruptcy prediction. The insights gained from this analysis, along with the proposed recommendations for improvement, provide a roadmap for ongoing refinement of these models. Continuous adaptation will ensure that these models remain effective in an ever-changing financial landscape.