# Project Report: Company Bankruptcy Prediction

**Presented by:**

- ***Yeshwanth Vemula (A20427054)*** - yvemula1@hawk.iit.edu
- ***Sundar Machani (*A20554747) – *smachani@hawk.iit.edu***

Table of Contents

# 1. Introduction

## 1.1 Background

Financial stability is pivotal for business success, necessitating accurate bankruptcy prediction. Conventional methods may miss subtle patterns, making machine learning crucial. This project applies advanced machine learning to enhance prediction accuracy, addressing the need for sophisticated financial analytics tools.

## 1.2 Objective

The project aims to leverage machine learning for precise bankruptcy prediction based on financial data, focusing on:

1. Diverse Models: Explore various machine learning models to understand their suitability for bankruptcy prediction.

2. Accurate Models: Develop reliable models with a balance between precision and recall.

3. Algorithmic Diversity: Evaluate Logistic Regression, Random Forest, XGBoost, KNeighbourClassifier, and Deep Learning (MLPClassifier) to identify the most effective approach.

4. Robustness and Generalization: Ensure models perform well on both training and test data for practical applicability.

5. Model Interpretability: Enhance transparency by interpreting model decisions, providing valuable insights into influential financial indicators.

In summary, this project comprehensively explores machine learning for bankruptcy prediction, aiming to deliver models that support informed decision-making in financial analysis.

# 2. Data

## 2.1 Dataset Description

The *Company Bankruptcy Prediction* dataset serves as the foundation for bankruptcy prediction, encompassing financial attributes with a binary target variable 'Bankrupt?'. Preprocessing addresses categorical variables, standardizes numerical features, and handles class imbalance using SMOTE.

## 2.2 Data Preprocessing

- Label Encoding: Categorical variables are label-encoded for machine learning model compatibility.
- Feature Scaling: Numerical features are standardized with StandardScaler for uniform model contribution.

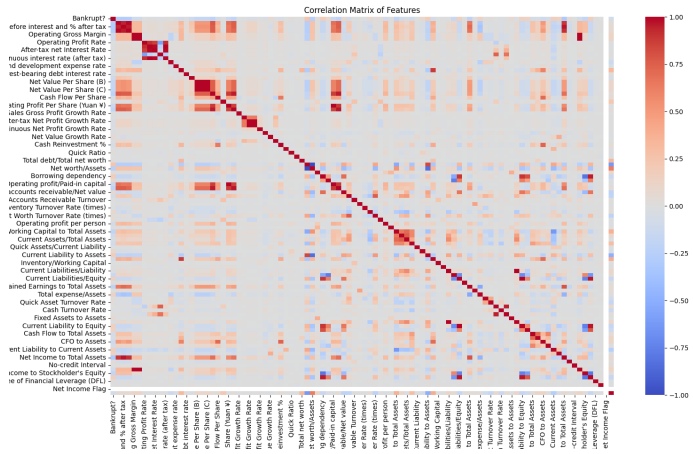
- Handling Class Imbalance with SMOTE: SMOTE addresses imbalanced 'Bankrupt?' classes.

## Attribute Information

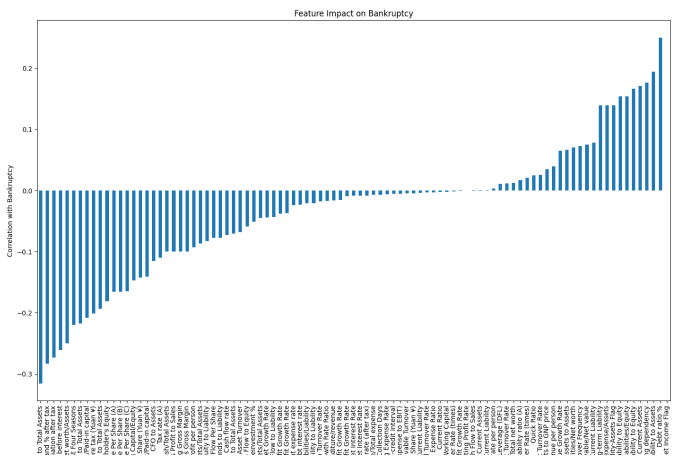
The dataset (X1 to X95) provides comprehensive financial features reflecting various aspects of a company's performance, including return ratios, profit margins, and leverage indicators. Understanding each feature is crucial for model interpretation.

## 2.3 Data Stability and Exploration

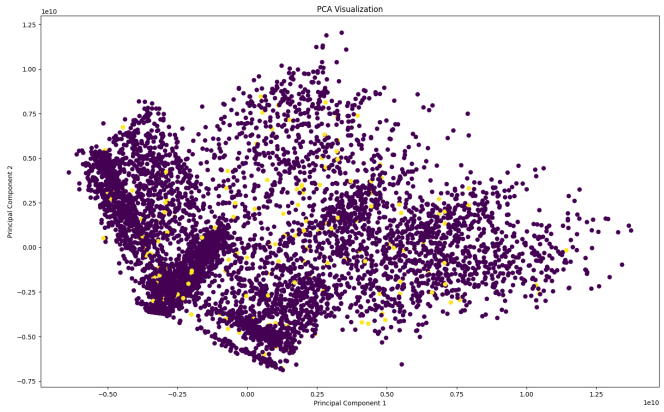
- Stability Distribution: 96.77% stable instances, 3.23% unstable/Bankrupt instances.
- Correlation Analysis: Heatmap explores relationships between features.



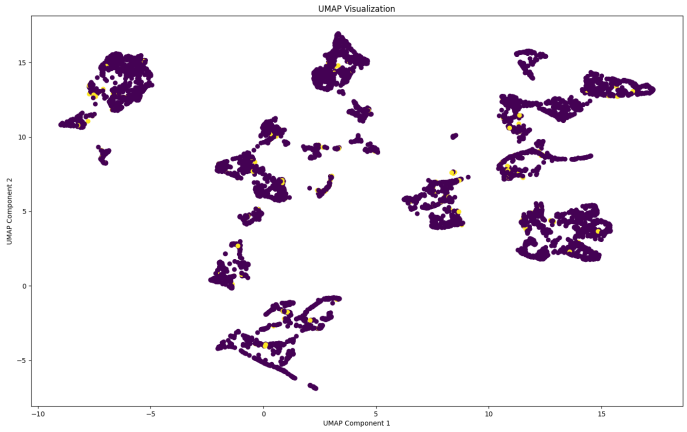

- Feature Impact on Bankruptcy: Bar chart visualizes the impact of features on bankruptcy.



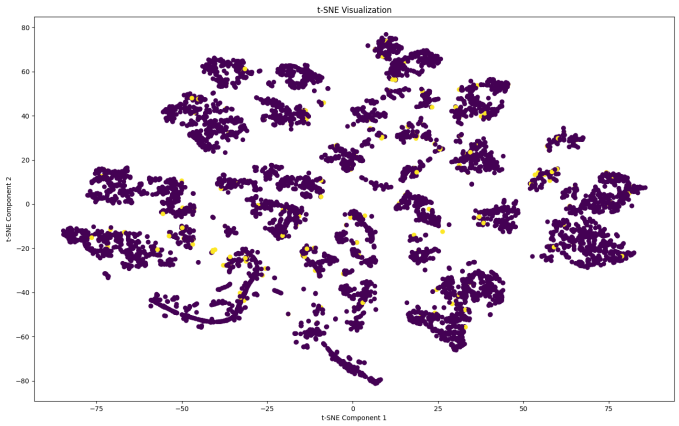

- **Dimensionality Reduction Techniques:**
- PCA Visualization: Reduces dimensionality to two components.




- UMAP Visualization: Provides insights into the data structure.




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




Observations:

- Initial insights gained from correlation matrix and feature impact analysis.
- Dimensionality reduction techniques offer diverse perspectives on data distribution.

- Visualizations serve as a foundation for further analysis and model building, guiding future investigation and experimentation.

# 3. Model Training

## 3.1 Logistic Regression (with Feature Selection)

Model Description
Logistic Regression with feature selection using another logistic regression model.

Results
Validation Accuracy: 89.32%
Cross-Validation Mean Accuracy: 89.31%
Test Accuracy: 89.32%
Precision, recall, and f1-score for classes 0 and 1.

## 3.2 Random Forest

Model Description
Random Forest classifier with 20 estimators for bankruptcy prediction.

Results
Validation Accuracy: 97.39%
Cross-Validation Mean Accuracy: 96.48%
Test Accuracy: 97.39%
Precision, recall, and f1-score for both classes.

## 3.3 XGBoost

Model Description
XGBoost, a gradient boosting algorithm, with specific configurations.

Results
Validation Accuracy: 97.88%
Cross-Validation Mean Accuracy: 97.44%
Test Accuracy: 97.88%
Precision, recall, and f1-score for both classes.

## 3.4 KNeighbourClassifier

Model Description
KNeighbourClassifier, a k-nearest neighbors algorithm, for bankruptcy prediction.

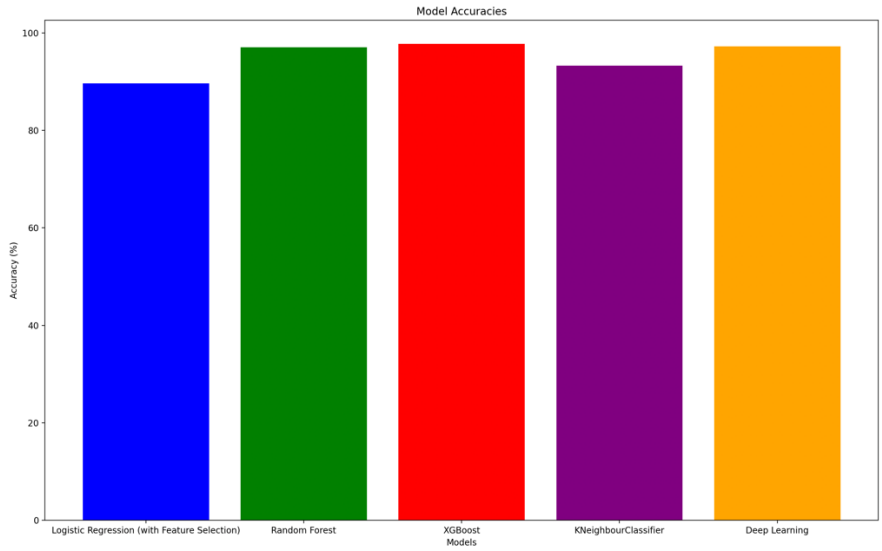
Results
Validation Accuracy: 93.75%
Cross-Validation Mean Accuracy: 93.41%
Test Accuracy: 93.75%

- Precision, recall, and f1-score for both classes.

### 3.5 Deep Learning (MLPClassifier)

Model Description
Multilayer Perceptron (MLP) Classifier for deep learning-based bankruptcy prediction.

Results
Validation Accuracy: 97.42%
Cross-Validation Mean Accuracy: 96.68%
Test Accuracy: 97.42%
Precision, recall, and f1-score for both classes.




Accuracy Plot of all models

## 4. Model Evaluation

### 4.1 Cross-Validation

Cross-validation, crucial for assessing model robustness, employs k-fold with 'k=5' for a balance between efficiency and reliable performance estimates. It helps identify issues like overfitting or underfitting by training on different subsets.

### 4.2 Performance Comparison

Models (Logistic Regression, Random Forest, XGBoost, KNeighbourClassifier, Deep Learning) are evaluated based on accuracy, precision, recall, and F1-score. Each model's strengths and weaknesses guide the selection of the most suitable one.

### 5. Potential Improvements

Identified Pitfalls
Convergence warnings observed for the Deep Learning model (MLPClassifier) may impact predictive performance.

Proposed Improvements
Deep Learning Model (MLPClassifier):
Increase Maximum Iterations: Adjust `max_iter` to address convergence warnings.
Fine-Tune Hyperparameters: Optimize parameters for better convergence.

General Recommendations:
Ensemble Methods: Explore stacking or blending for improved accuracy.
Feature Engineering: Experiment with creating or transforming features.

### 5.3 Additional Experiments

Algorithm Variation:
Experiment with alternative algorithms or variations for a more comprehensive understanding.

Feature Selection Methods:
Evaluate different feature selection methods beyond Logistic Regression.

These experiments aim to enhance predictive performance and provide a deeper understanding of the dataset and model behavior.

## 6. Code and Dataset

Code repository : https://github.com/sundarmachani/CS-584-machine-Learning-Final-Project

Kaggle Dataset used : https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction/data

## 7. Conclusion

This project focused on leveraging machine learning for bankruptcy prediction, yielding significant insights into company financial health. Key highlights include:

1. Dataset Exploration:
'Company Bankruptcy Prediction' provided a comprehensive set of financial attributes.
Preprocessing steps involved label encoding, feature scaling, and SMOTE to address class imbalance.

2. Model Development:
Explored algorithms: Logistic Regression, Random Forest, XGBoost, KNeighbourClassifier, and Deep Learning.
Logistic Regression facilitated feature selection.

3. Model Evaluation:
Utilized a robust cross-validation strategy for thorough evaluation.
Comparative analysis revealed algorithm strengths and weaknesses.

Final Thoughts:

This project establishes a solid framework for machine learning in bankruptcy prediction. Insights gained, coupled with recommendations, pave the way for ongoing model refinement. Continuous adaptation ensures these models remain effective in dynamically evolving financial landscapes.