Project Title: Corporate Bankruptcy Prediction and Financial Ratio Clustering

Submitted by:

Prathamesh Kulkarni(122B1B147) , Amit Deshmukh (121B1B050)
Shivam Hucche(121B1B069) , Omkar Kudke(122B1B146)

Submitted to: Professor Tamboli Mubin , Computer department , Pccoe

Date: 4/4/25

Table of Contents

List of Figures and Tables

Tables

1. Introduction

Background of the Project

Corporate bankruptcy significantly impacts stakeholders and economic stability. Early bankruptcy prediction can help mitigate negative effects, guiding investors, creditors, and employees. Financial ratios have long been used to predict bankruptcy, but traditional methods face limitations in adapting to modern financial dynamics.

Problem Statement

Predicting corporate bankruptcy using financial ratios is a challenging task due to issues like data imbalance, economic fluctuations, and firm heterogeneity. There is a need for an advanced prediction model that integrates machine learning and unsupervised clustering to improve the prediction accuracy.

Objective of the Project

To predict corporate bankruptcy using a hybrid approach that combines machine learning classifiers with unsupervised clustering techniques, and to evaluate the effect of adding clustering-derived features on model performance.

Scope of the Project

This project explores the application of machine learning techniques, such as logistic regression, decision trees, random forests, SVM, k-NN, and XGBoost, along with clustering methods like K-Means, to enhance corporate bankruptcy prediction.

2. Literature Review / Existing System

Overview of Existing Solutions

Early approaches such as Altman's Z-score and Ohlson's logistic regression laid the foundation for bankruptcy prediction. These models relied on linear statistical methods but were limited in adaptability to complex financial data.

Limitations of the Existing System

Existing systems, while foundational, struggle with non-linearity in financial data and class imbalance. Moreover, they fail to incorporate unsupervised insights to better segment firms based on financial behaviors.

Hybrid Model Approaches

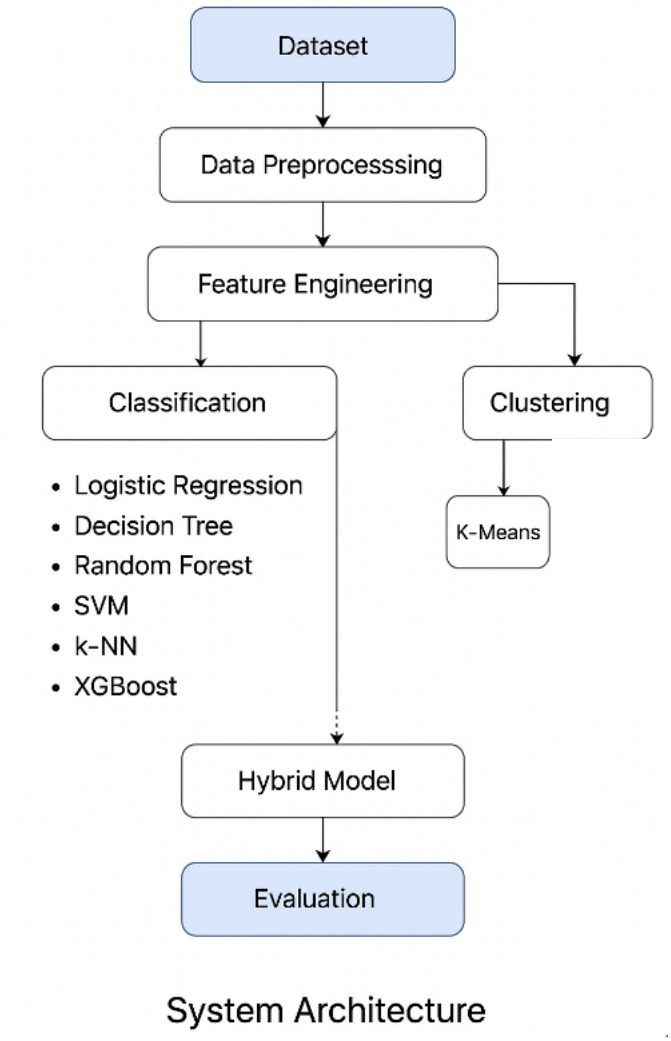
Recent studies have explored hybrid models that integrate clustering and classification, which address class imbalance and provide a more granular understanding of the firms' financial health.

3. Proposed System

Explanation of the System

The system developed uses multiple classification algorithms combined with unsupervised clustering to improve bankruptcy prediction. The financial ratios are processed, and K-Means clustering is applied to label firms into different financial risk categories. These labels are then integrated into classifiers to evaluate if they improve model performance, especially in terms of recall for bankrupt firms.

System Architecture



System Architecture

The system is composed of a data preprocessing module, a classification module (with multiple machine learning models), and a clustering module. The models are then trained on financial ratios, with clustering features included in the final predictions.

Features and Advantages over Existing System

• Integration of clustering labels to improve model accuracy and recall for bankrupt firms.

• Use of advanced machine learning models like XGBoost and Random Forest to capture non-linear patterns.

• Higher precision in identifying bankrupt firms due to improved data segmentation.

4. Implementation & Results

Technologies and Tools Used

• Programming Languages: Python
• Libraries: Scikit-learn, Pandas, Numpy, Matplotlib
• Machine Learning Algorithms: Logistic Regression, Decision Tree, Random Forest, SVM, k-NN, XGBoost
• Clustering Techniques: K-Means, Hierarchical Clustering
• Performance Metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix

Dataset Information

The dataset includes various financial ratios for companies, with labels indicating whether a company went bankrupt. Key ratios include liquidity, leverage, profitability, and activity ratios.

Performance Metrics Used

• Accuracy: Measures overall prediction correctness.
• Precision: Assesses the proportion of true positives among all predicted positives.
• Recall: Measures the ability to identify actual bankrupt companies.
• F1-Score: Combines precision and recall for a balanced measure.
• Confusion Matrix: Provides detailed performance analysis, including false positives and false negatives.

Description of Major Modules

• Data Preprocessing: Cleans and standardizes the data for model input.
• Classification: Trains various classifiers and evaluates their performance.
• Clustering: Applies clustering to segment firms based on financial similarity.
• Hybrid Model: Enhances classification by incorporating clustering labels as features.

Comparative Results

Classifier	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Logistic Regression	0.9322	0.5385	0.0066	0.0130	[[14663, 6], [1061, 7]]
Decision Tree	0.8923	0.2243	0.2388	0.2313	[[13787, 882], [813, 255]]
Random Forest	0.9363	0.9333	0.0655	0.1225	[[14664, 5], [998, 70]]
SVM	0.9325	1.0000	0.0056	0.0112	[[14669, 0], [1062, 6]]
k-Nearest Neighbors	0.9402	0.6721	0.2322	0.3452	[[14548, 121], [820, 248]]
XGBoost	0.9361	0.6962	0.1030	0.1794	[[14621, 48], [958, 110]]

• Table 1: Baseline models without clustering show decent performance, with k-NN having the highest accuracy of 0.9402.

Classifier	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Hybrid Logistic Regression	0.9322	0.5385	0.0066	0.0130	[[14663, 6], [1061, 7]]
Hybrid Decision Tree	0.8922	0.2133	0.2191	0.2162	[[13806, 863], [834, 234]]
Hybrid Random Forest	0.9361	0.9437	0.0627	0.1176	[[14665, 4], [1001, 67]]
Hybrid SVM	0.9325	1.0000	0.0056	0.0112	[[14669, 0], [1062, 6]]
Hybrid k-Nearest Neighbors	0.9401	0.6721	0.2303	0.3431	[[14549, 120], [822, 246]]
Hybrid XGBoost	0.9361	0.6962	0.1030	0.1794	[[14621, 48], [958, 110]]

Table 2: Hybrid models (with clustering labels) show modest improvement in recall, particularly for identifying bankrupt firms.

5. Conclusion and Future Scope

Summary of the Project

This research successfully integrates clustering with supervised classification to enhance corporate bankruptcy prediction. The hybrid approach improves recall for bankrupt firms, offering better predictive power in imbalanced datasets. The k-NN classifier showed the highest accuracy, while Random Forest performed well in precision.

Future Scope

- Further optimization of clustering techniques for more distinct groupings.
- Application of deep learning models for bankruptcy prediction.
- Exploration of other clustering algorithms such as DBSCAN or Gaussian Mixture Models for improved segmentation.

6. References

1. E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968.

2. J. A. Ohlson, “Financial ratios and the probabilistic prediction of bankruptcy,” Journal of Accounting Research, vol. 18, no. 1, pp. 109–131, 1980.

3. C.-F. Tsai, “Two-stage hybrid learning techniques for bankruptcy prediction,” Statistical Analysis and Data Mining, vol. 13, no. 6, pp. 565–572, 2020.

4. D. Liang et al., “Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study,” European Journal of Operational Research, vol. 252, no. 2, pp. 561–572, 2016.