Company Bankruptcy Prediction:

A Hybrid Framework Leveraging Combined Probabilities from Gaussian and Neural Networks

Mainak Ch Rishitha
mainak23@iitk.ac.in rishitha23@iitk.ac.in
230619 230311

1. INTRODUCTION

Predicting a company's bankruptcy has become a critical task in today's uncertain economic conditions. This project aims to build a robust machine learning model that can predict bankruptcy by exploring a range of factors which indicate the financial status of the company.

This report is structured as follows: we begin with Exploratory Data Analysis (EDA) to examine the dataset and uncover key patterns and correlations; next, we discuss Data Preprocessing, detailing how the data is cleaned, normalized, and balanced; then, we describe the Model Architecture, outlining the machine learning models used for bankruptcy prediction and the rationale behind their design; and finally, we evaluate the models using performance metrics such as accuracy, precision, recall, and F1 score to assess their practical effectiveness.

2. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) and feature engineering involves several key steps to understand and preprocess the dataset effectively. The primary goal was to understand the data's characteristics, identify key features, and prepare the dataset for model training.

The dataset used for bankruptcy prediction comprises 7027 rows, each corresponding to a company, and 66 financial features. The target variable distribution revealed a significant class imbalance, with 6756 non-bankrupt companies (97.2%) and 271 bankrupt companies (2.8%). This imbalance was considered in subsequent model development.

A preliminary analysis revealed extreme values were identified in several numerical features, potentially indicating data errors. Features with over 800 erroneous entries were discarded, while features with fewer than 200 errors underwent median imputation.

To mitigate multicollinearity and enhance model interpretability, a feature correlation analysis was conducted. Features with a Pearson correlation coefficient exceeding 0.9 were examined, and the one with the weaker correlation to the target variable was removed. This process reduced

the number of features from 63 to 50. A detailed correlation heatmap (Fig. 1) illustrates the relationships among retained features.

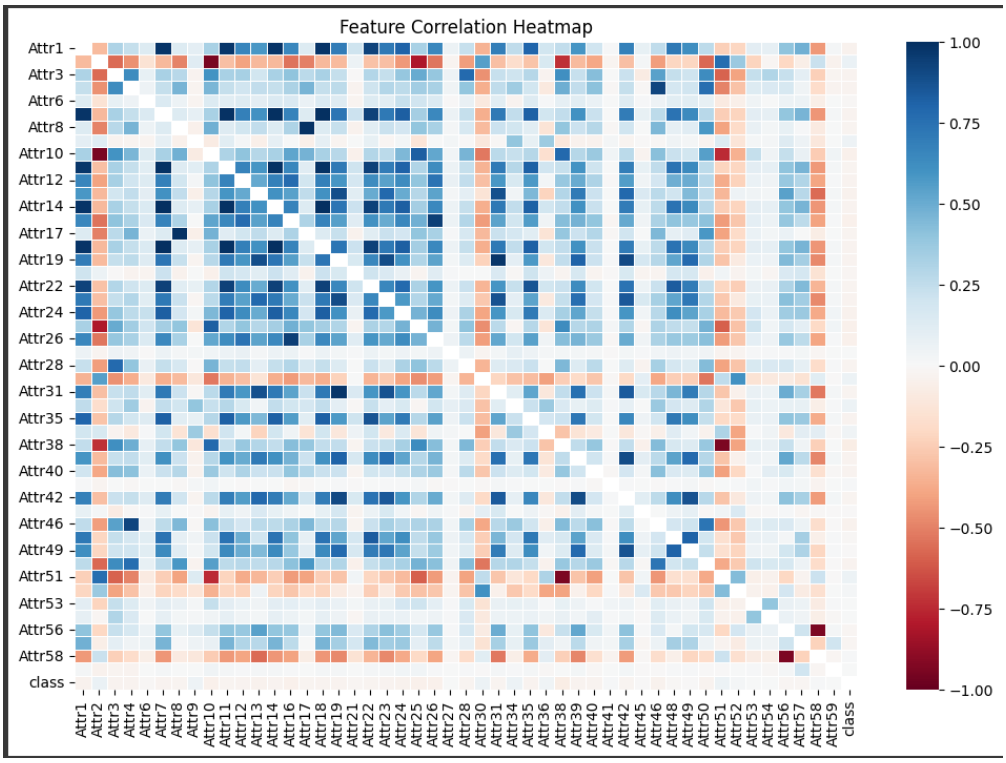

Figure 1: *Feature Correlation Heatmap of 62 features*

3. DATA PREPROCESSING

4. Analysis of Variance (ANOVA):

Analysis of Variance (ANOVA) was then used to select the most relevant features for predicting bankruptcy. ANOVA F-scores were calculated for each feature to assess its significance.

$$\text{F score} = \frac{\text{Variance between groups}}{\text{Variance within groups}} \tag{1}$$

Where:

$$\text{Variance between groups} = \sum_{i=1}^{k} n_i(\mu_i - \mu)^2$$

$$\text{Variance within groups} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2$$

A high F-score signifies substantial variation in feature values across target variable classes. The `SelectKBest` method, utilizing the `f_classif` scoring function, was employed to identify the top 30 features. These selected features ensured that the model concentrated on those exhibiting the most significant differences in behavior between the two classes.

[12pt]article graphicx amsmath hyperref xcolor

Hybrid Bankruptcy Prediction using Deep Learning and Probabilistic Modeling Your Name Teammate Name

B. OVERSAMPLING USING SMOTE

The dataset exhibited a significant class imbalance, with 6,756 non-bankrupt (96.1%) and 271 bankrupt (3.9%) companies. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data.

Following an 80-20 train-test split, the training set contained 5,405 non-bankrupt companies and 216 bankrupt companies. SMOTE was used to generate synthetic samples for the minority class, balancing the training set to 5,405 instances in each class. This oversampling was restricted to the training data to prevent biasing the test set. SMOTE operates by selecting a minority class sample, identifying its k-nearest neighbors, and generating a new synthetic data point through linear interpolation between the selected sample and one of its neighbors, increasing the representation of the minority class.

C. STANDARDISATION

To ensure uniform feature scaling, `StandardScaler` was applied, transforming all features to have a mean of 0 and a standard deviation of 1. This prevents dominance by features with larger magnitudes.

4. MODEL ARCHITECTURE

We developed a hybrid bankruptcy prediction model by ensembling a Deep Neural Network (DNN) and a Gaussian Naive Bayes (GNB) classifier. The objective was to leverage the probabilistic nature of GNB alongside the deep feature learning capabilities of DNN to improve performance.

The first model, DNN, was developed to capture complex, non-linear relationships between financial indicators. It consists of an input layer, three hidden layers, and an output layer. The input layer receives the selected 30 features. The first hidden layer comprises 256 neurons, utilizing the ReLU activation function, batch normalization, and dropout (50%) to prevent overfitting. The

second hidden layer has 128 neurons, also incorporating batch normalization and dropout (50%). The third hidden layer refines the feature representations further with 64 neurons and a reduced dropout (40%). The output layer contains a single neuron with a sigmoid activation function, outputting a probability score for bankruptcy classification.

The DNN model was compiled with the Adam optimizer (learning rate = 0.0005) and binary cross-entropy loss function. It was trained for 200 epochs with a batch size of 64 and a 20% validation split.

GNB applies Bayes' theorem, assuming feature independence and normal distribution given the class label. It computes the posterior probability for each class using the prior probability and the likelihood, where the likelihood of continuous features is modeled using a normal distribution. The model predicts the class with the highest posterior probability.

Initially, employing only the DNN with feature selection yielded a maximum F1-score of 0.46. By integrating GNB into the ensemble, the model effectively leveraged probabilistic classification, improving the F1-score to 0.51.

To leverage both models, we applied an ensemble approach using soft voting. The probability outputs from the DNN and GNB models were averaged to compute the final bankruptcy probability. Instead of using the default classification threshold of 0.5, we fine-tuned the threshold by evaluating F1-scores across multiple threshold values (between 0.30–0.60). The threshold (0.45) that maximized the F1-score was selected for final predictions.

5. Performance Metrics of the Model

The Gaussian Naive Bayes and DNN ensemble model reached 97.23% accuracy on the test set. Other result metrics in the Classification Report (Fig. 3) along with the Confusion Matrix (Fig. 2) are shown below:

6. Conclusion

This study introduced a hybrid bankruptcy prediction framework that integrates a Deep Neural Network (DNN) with a Gaussian Naïve Bayes (GNB) classifier. Rigorous exploratory data analysis and feature engineering—including SMOTE for class imbalance, ANOVA-based feature selection, and standardization—ensured robust data preprocessing.

The ensemble, which combines deep feature learning and probabilistic inference through fine-tuning the decision threshold and soft voting for both advanced models, achieved a test accuracy of 97.23% and an F1-Score of 0.51 while maintaining a precision–recall trade-off. This result is noteworthy given the data with extreme imbalance (class ratio of 25:1). These results underscore

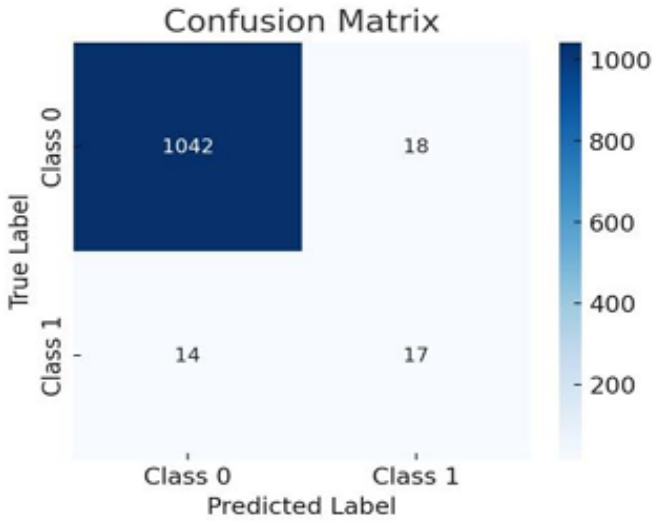

Figure 2: *Figure 2: Confusion Matrix*

Classification Report:				
	precision	recall	f1-score	support
0	0.99	0.98	0.98	1060
1	0.49	0.55	0.52	31
accuracy			0.97	1091
macro avg	0.74	0.77	0.75	1091
weighted avg	0.97	0.97	0.97	1091

Figure 3: Classification Report

Figure 3: *Figure 3: Classification Report*

the framework's potential for accurate bankruptcy prediction, laying the groundwork for future enhancements in feature selection and ensemble strategies.

REFERENCES

[1] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pp. 109–131.

[2] Kong, Y. (2009). Semi-supervised learning and its application research [D]. Wuxi: Jiangnan University, pp. 33–39. *Advances in Intelligent Systems Research*, Vol. 168399.

[3] Dong, L., Sui, P., Sun, P., & Li, Y. (2016). A new naive Bayesian algorithm based on semi-supervised learning. *Journal of Jilin University (Engineering Science Edition)*, 46(3), pp. 884–889.

[4] Abdullah, S. A., & Al-Ashoor, A. (2010). An artificial deep neural network for the binary classification of network traffic. *Journal of Advanced Computer Science and Applications*, 11(1).

[5] Bi, Z., Han, Y., Huang, C., & Wang, M. (2016). Gaussian Naive Bayesian data classification model based on clustering algorithm.