Submitted By: *Waleed Malik*

Job Title: *AI/ML Engineer*

Company Name: *Joblogic*

Assessment Report

Objective

To develop a machine learning model that can predict whether a company is likely to go bankrupt based on its financial indicators collected over 5 years. The task is framed as a binary classification problem (0 = non-bankrupt, 1 = bankrupt).

Problem-Solving Approach

1. Understanding the Data The dataset contains 64 numerical financial features per company for each year. After merging all yearly datasets, we noticed:

Class imbalance: Very few bankrupt companies compared to non-bankrupt ones

Missing data: Some features had significant missing values

2. Data Preprocessing Dropped columns with more than 40% missing values

Imputed remaining missing values using the median (to handle outliers better)

Standardized all features using StandardScaler to normalize the range

Handled class imbalance using SMOTE, which oversamples the minority class (bankrupt) by synthetically generating new examples

3. Exploratory Data Analysis Histograms showed that many features were skewed, which is common in financial data

Correlation heatmap helped understand redundancy and potential multicollinearity

Applied PCA to visualize class separability in a 2D space

Model Selection

Logistic Regression algorithm for linear and binary classification problems. can be readily generalized to multiclass settings, which is known as multinomial logistic regression or softmax regression.

Random Forest A random forest can be considered as an ensemble of decision trees. The idea behind a random forest is to average multiple (deep) decision trees that individually suffer from high variance to build a more robust model that has a better generalization performance and is less susceptible to overfitting.

XGBoost Often top performer for structured/tabular data; handles imbalance well. It is essentially a computationally efficient implementation of the original gradient boost algorithm.

Evaluation

Train-Test Split (80/20 stratified) to ensure class balance

Performance Metrics:

Accuracy

Precision, Recall, F1-score (to handle imbalance)

ROC-AUC (for overall class separability)

Cross-Validation: Used StratifiedKFold to validate model robustness (5 folds)

Model Results

Logistic Regression gave good baseline performance but struggled with recall missed some bankruptcies

Random Forest performed better overall, especially in identifying bankrupt companies

XGBoost had the best ROC-AUC and balanced precision/recall, making it the top choice

---

Conclusion

For this task, XGBoost was the most effective model due to its ability to handle imbalance, capture complex patterns, and deliver strong predictive performance. Random Forest was a close second and also helped interpret which financial features were most important.