PREDICTING CORPORATE BANKRUPTCY

STUDIO SHODWE

PROJECT OVERVIEW

We are trying to predict whether a company will go bankrupt. This is important because it helps investors, banks, and businesses avoid financial losses. If we can predict bankruptcy accurately, companies can take early action to prevent failure, and banks can make better lending decisions. We aim to develop a machine learning model that can classify companies as either bankrupt or not bankrupt based on financial indicators.

STUDIO SHODWE

OBJECTIVES

1. Data Collection and Preprocessing
2. Model Training
3. Model Evaluation and Feature importance

PREPROCESSING

Data Overview

- Dataset: The dataset contains 6819 rows and 96 features, including financial indicators like profitability, debt ratios, and asset management.
- Missing Values: No missing data, all columns are complete.
- Target Variable: 'Bankrupt?' (0 for non-bankrupt, 1 for bankrupt).

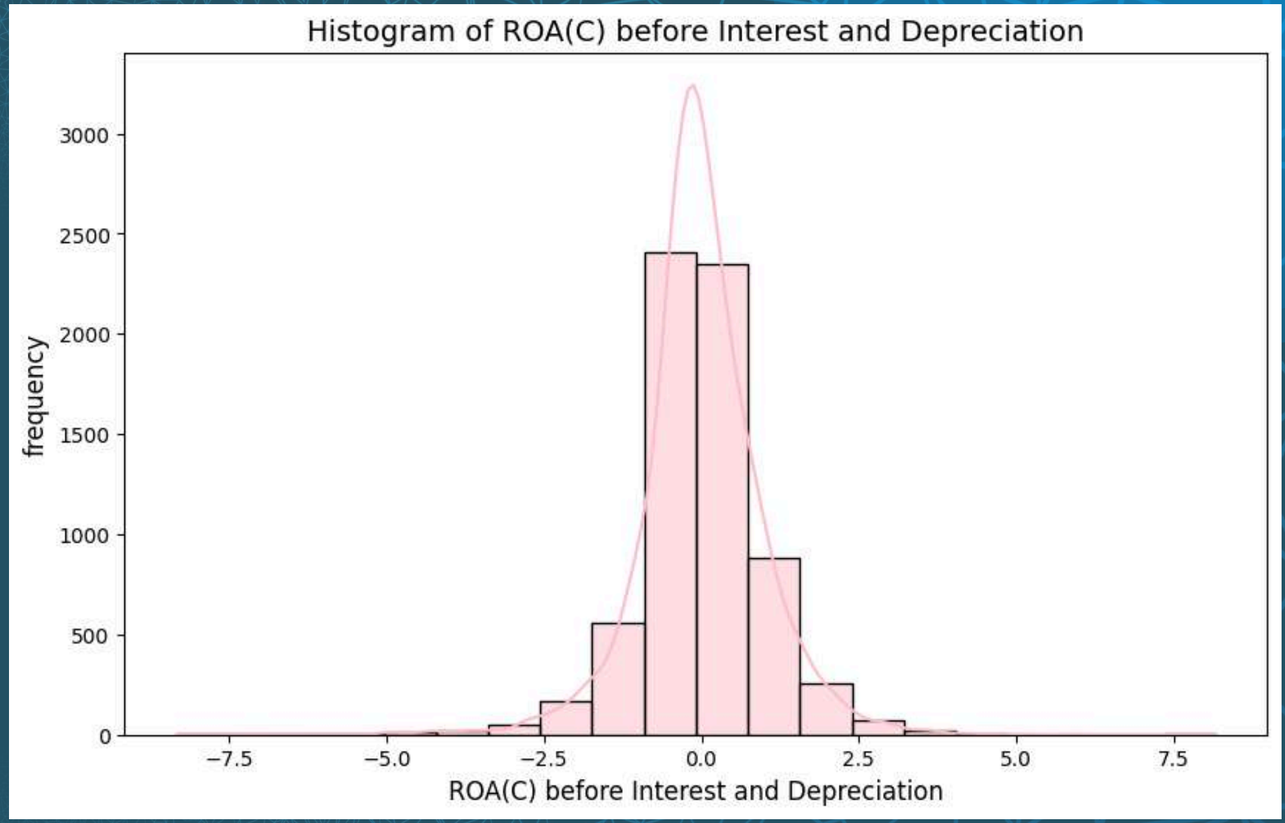
Data Cleaning

- Removing duplicates: No duplicate entries.
- Feature Scaling: Standardized the features using StandardScaler to bring all data to the same scale.
- Train-Test Split: Split data into 80% training and 20% testing.

EDA TECHNIQUES

Histogram Interpretation

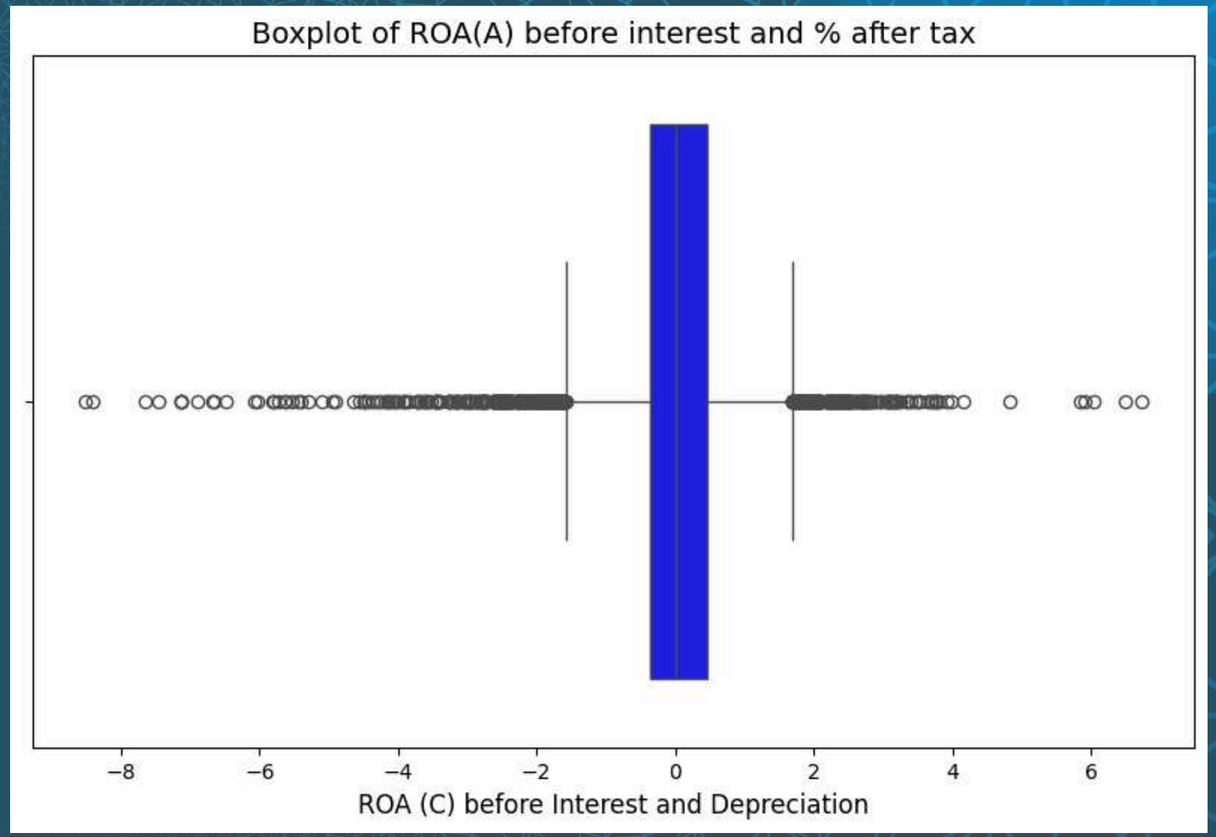
A histogram shows how data is distributed across different ranges. It displays the frequency of data points in intervals, helping us see patterns, trends, and outliers easily. The x-axis shows the data values, and the y-axis shows how often those values occur.



EDA TECHNIQUES

Boxplot Interpretation

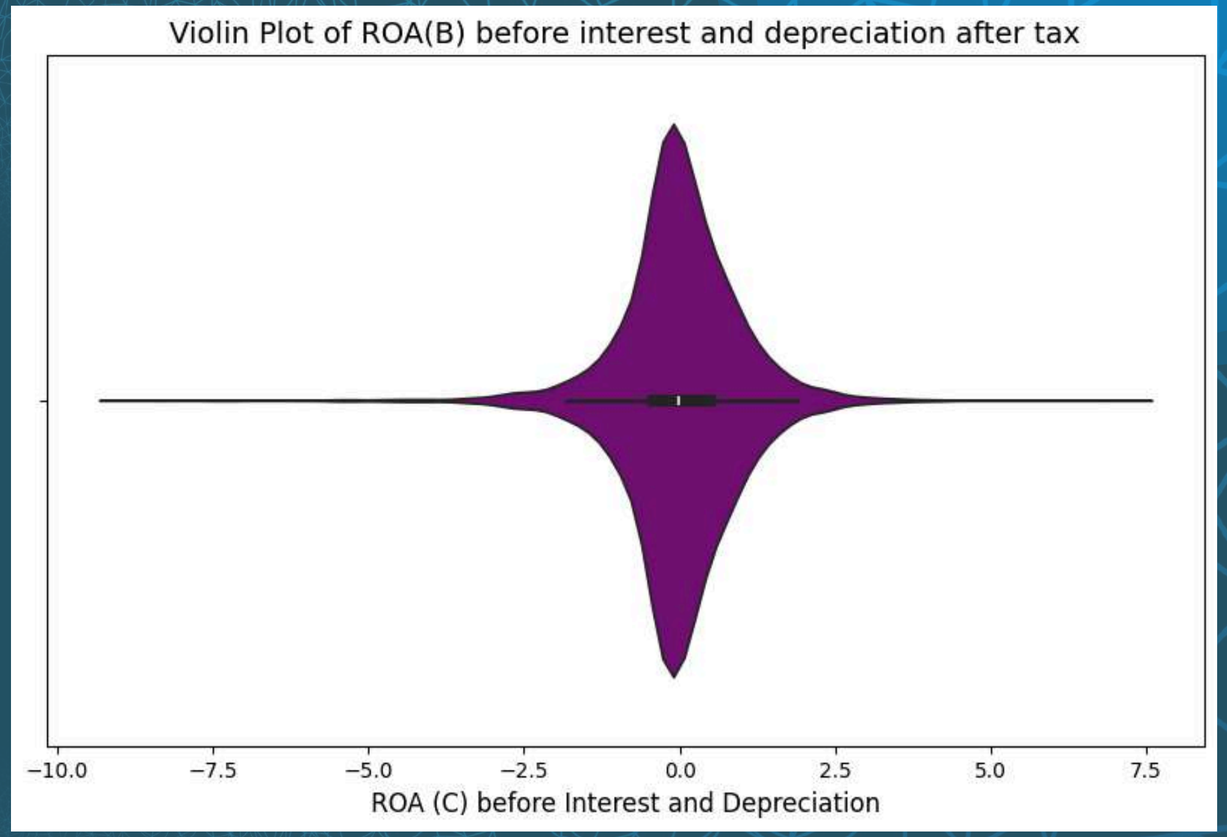
A boxplot shows the distribution of data, highlighting the median, quartiles, and any outliers. It uses a box to represent the interquartile range, showing the range of the data. Points outside the box are considered outliers.



EDA TECHNIQUES

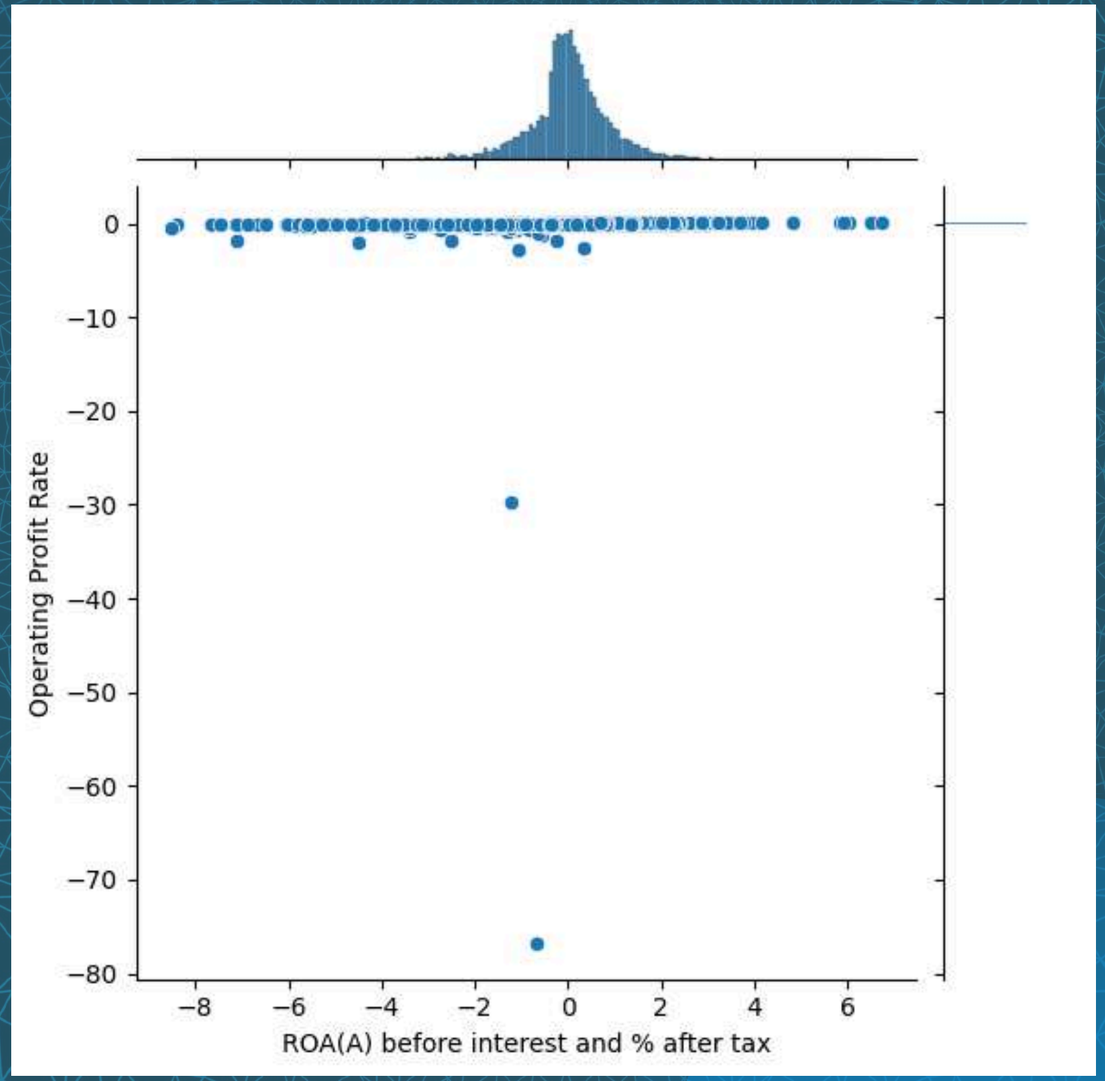
Violin Plot Interpretation

A violin plot combines a boxplot and a density plot, showing the distribution and density of data at different values. The wider sections of the "violin" represent areas with more data, while the line in the center indicates the median.



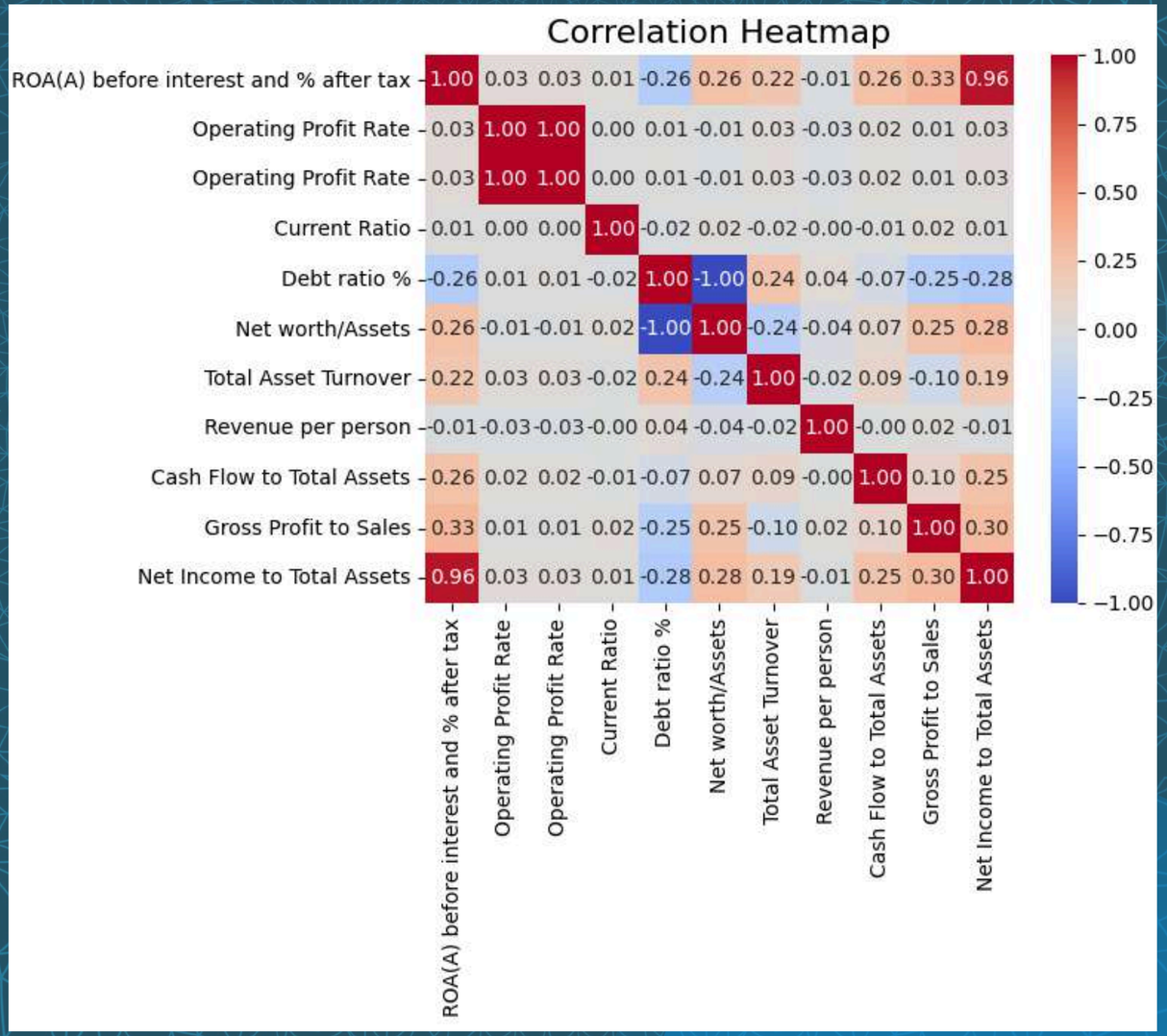
EDA TECHNIQUES

Joint Plot



EDA TECHNIQUES

Heatmap



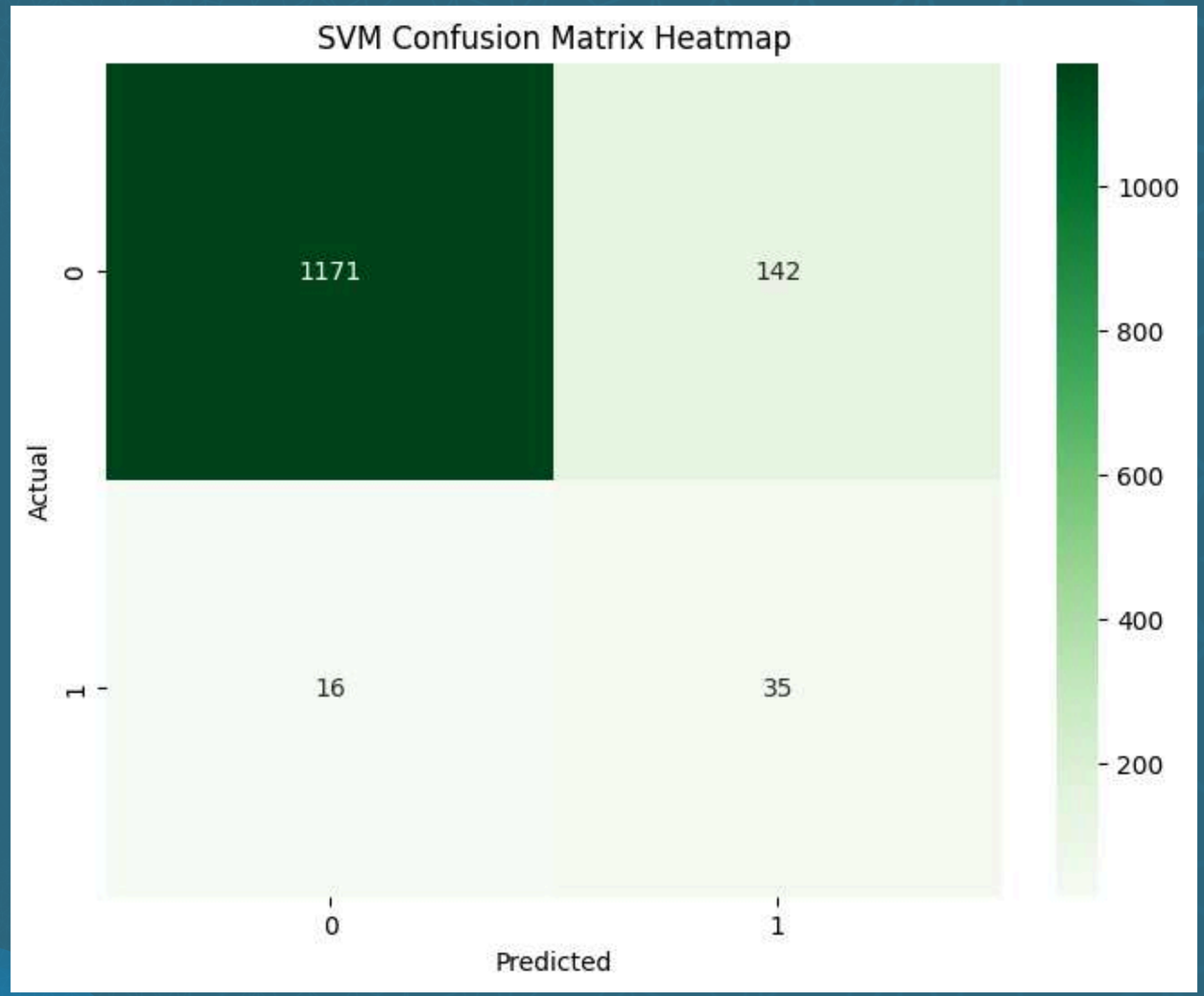
SVM model Classifier with Recursive Feature Elimination (RFE)

SVM MODEL CLASSIFIER

SVM Classifier: Trained an SVM classifier with an RBF kernel.

Accuracy: 88.42%

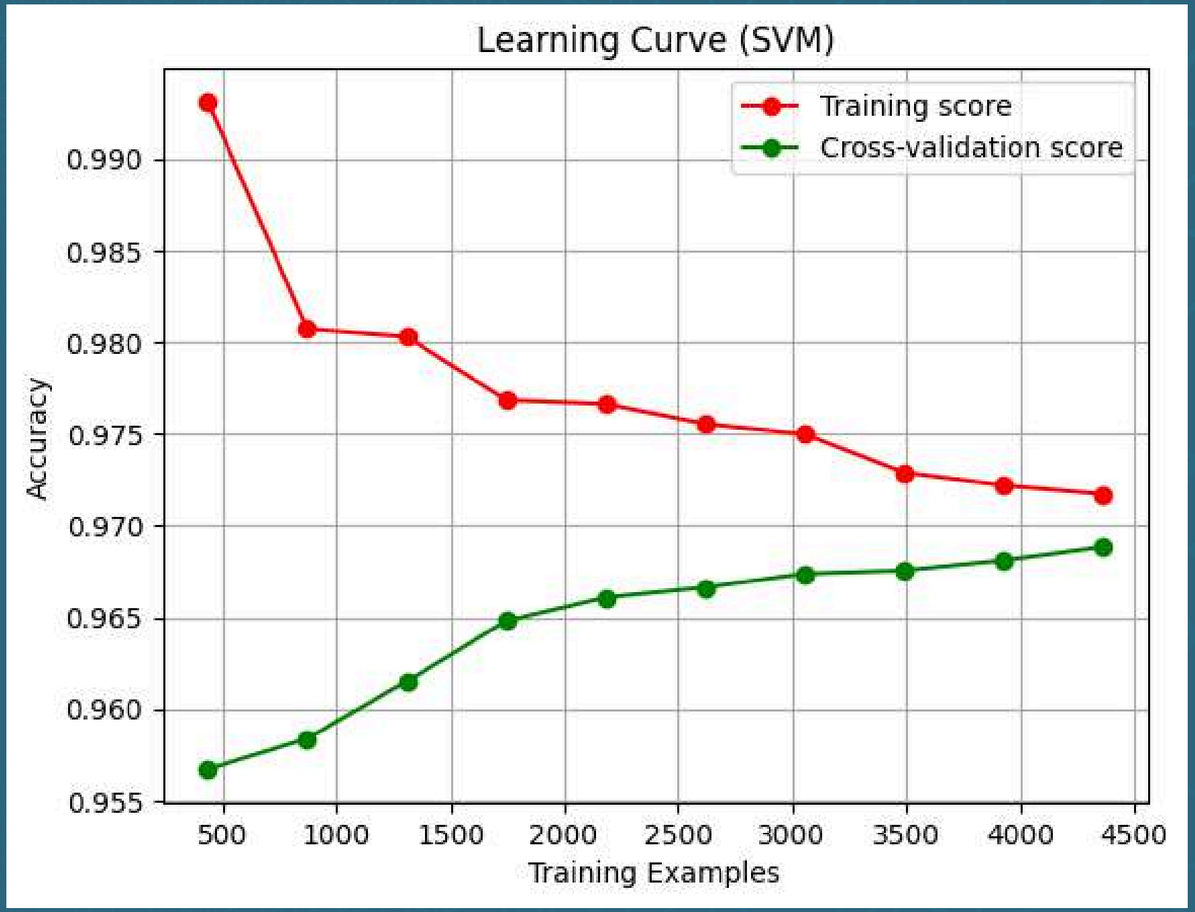
- **Confusion Matrix:**
  - **True Negatives: 1171**
  - **False Positives: 142**
  - **False Negatives: 16**
  - **True Positives: 35**



SVM MODEL CLASSIFIER

. Learning Curve

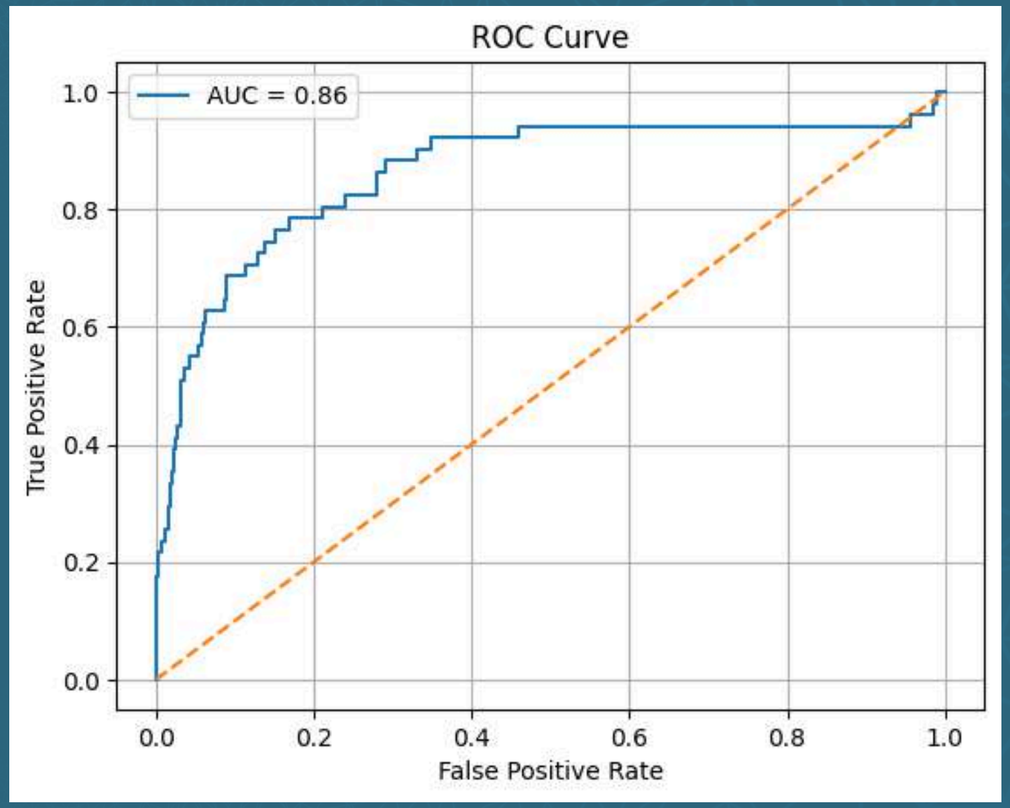
- A learning curve shows how well a model is learning over time.
- The x-axis shows the number of training samples.
- The y-axis shows the model's accuracy.
- Two lines are shown: one for training performance and one for validation performance.
- If the validation line stays low, it may mean underfitting.
- If the gap between training and validation is big, it may mean overfitting.



SVM MODEL CLASSIFIER

ROC Curve

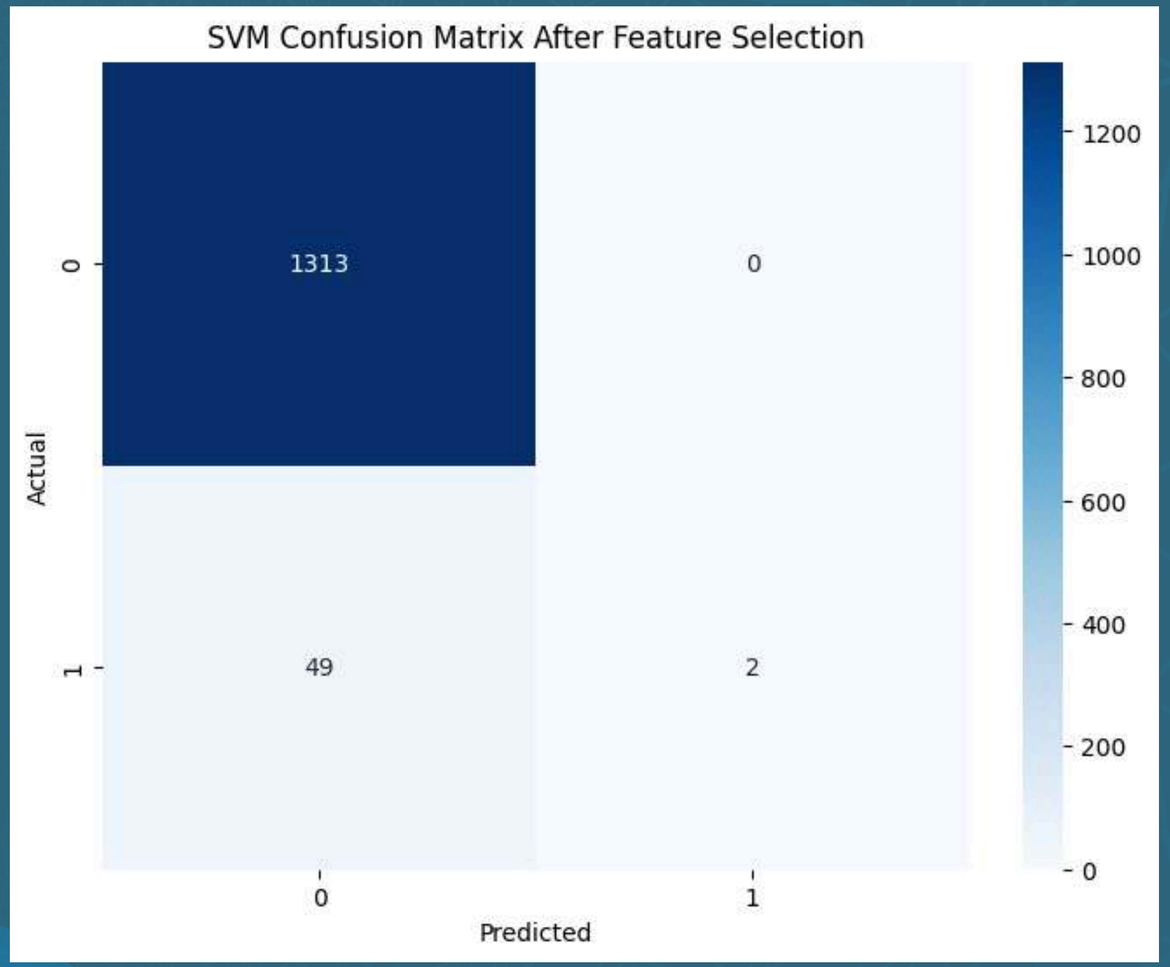
- The ROC curve is a graph that shows how well a classification model can separate between classes.
- It plots True Positive Rate (TPR) vs. False Positive Rate (FPR).
- The Area Under the Curve (AUC) tells how good the model is:
- AUC = 1: The model is good at telling the difference between positive and negative cases.
- AUC = 0.5: The model isn't learning anything useful and is just guessing randomly.
- AUC < 0.5: The model struggles to tell the difference between positive and negative classes



SVM MODEL CLASSIFIER

Model Performance After Feature Selection:

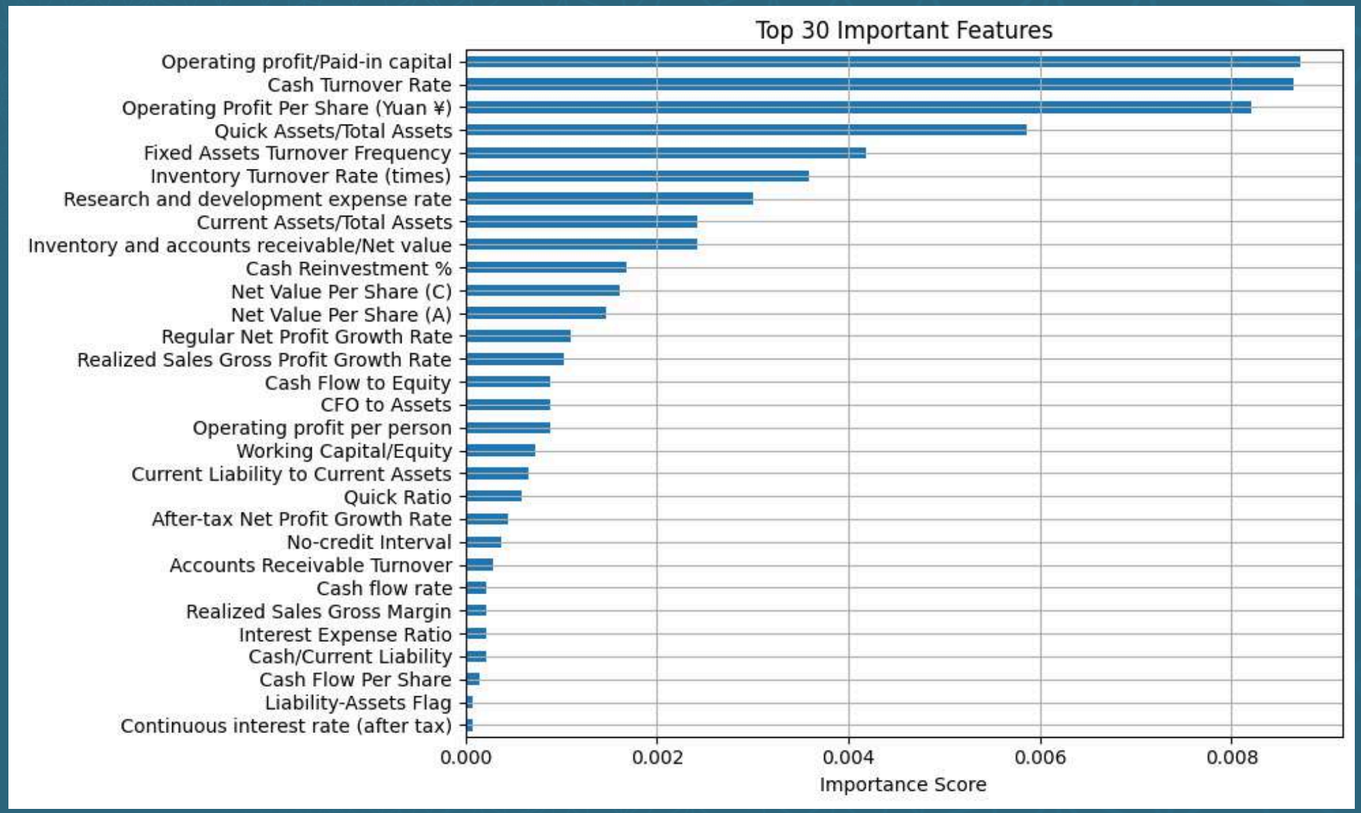
- *Final Model: After feature selection, the SVM model achieved 96.4% accuracy.*
- *Confusion Matrix: Improved performance with higher precision and recall for the bankrupt class.*
- *Final Classification Report: Enhanced model performance on classifying bankrupt companies.*



SVM MODEL CLASSIFIER

Permutation importance

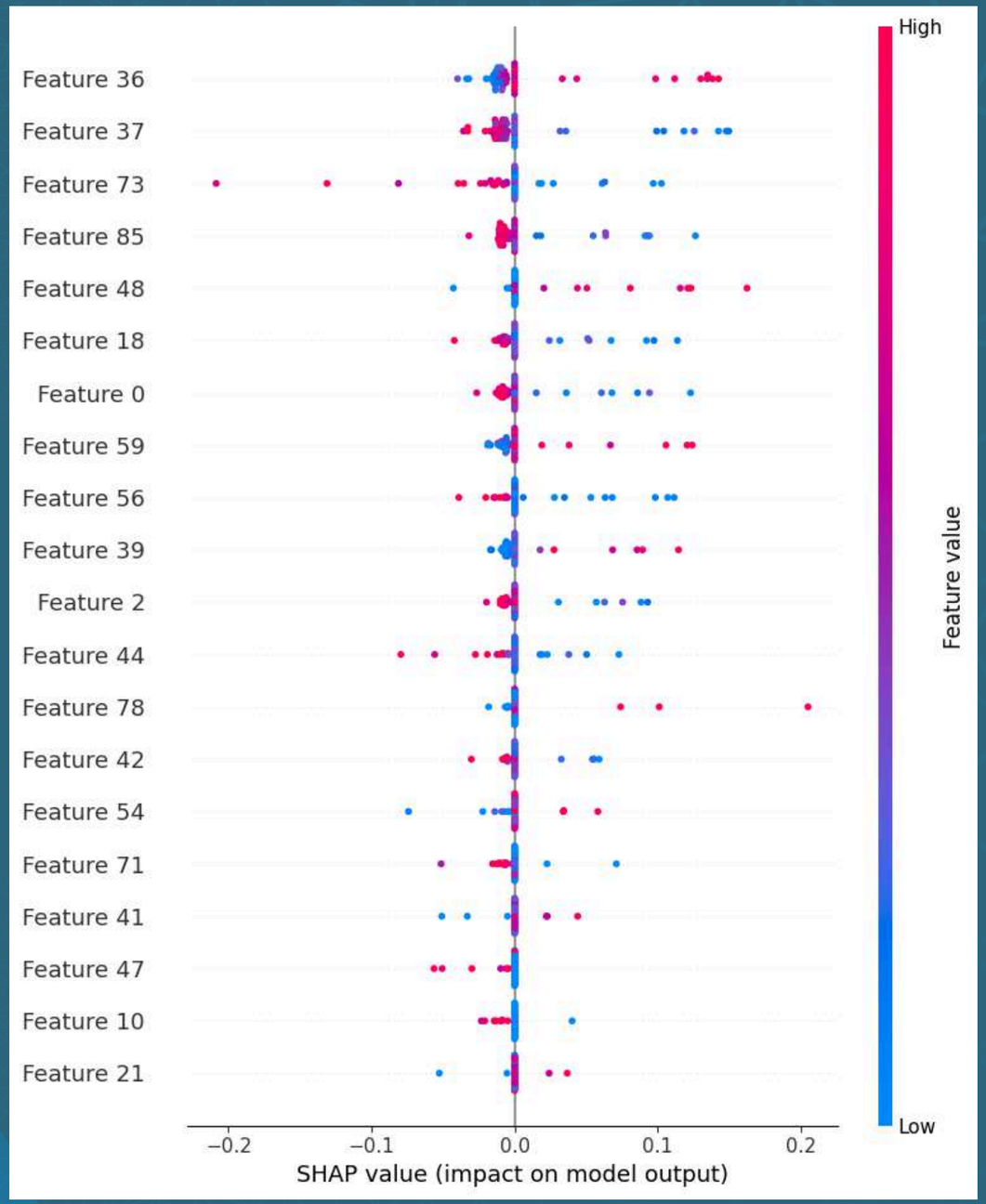
- Permutation importance tells us how much each feature important in the model.
- It works by randomly shuffling one feature at a time and checking how much the model's performance drops.
- If the performance drops a lot, that feature is important.
- If there's no change or improvement, that feature is not important.



SVM MODEL CLASSIFIER

SHAP

- SHAP. It works based on game theory: Imagine features are like players in a game working together to make a prediction. SHAP tells us how much each feature contributes. How SHAP works: Treat each feature value as a player. Prediction is the game's final output. Calculate how much the outcome changes with and without each feature, and this is called the Shapley value. Average these changes across all combinations to get the final importance



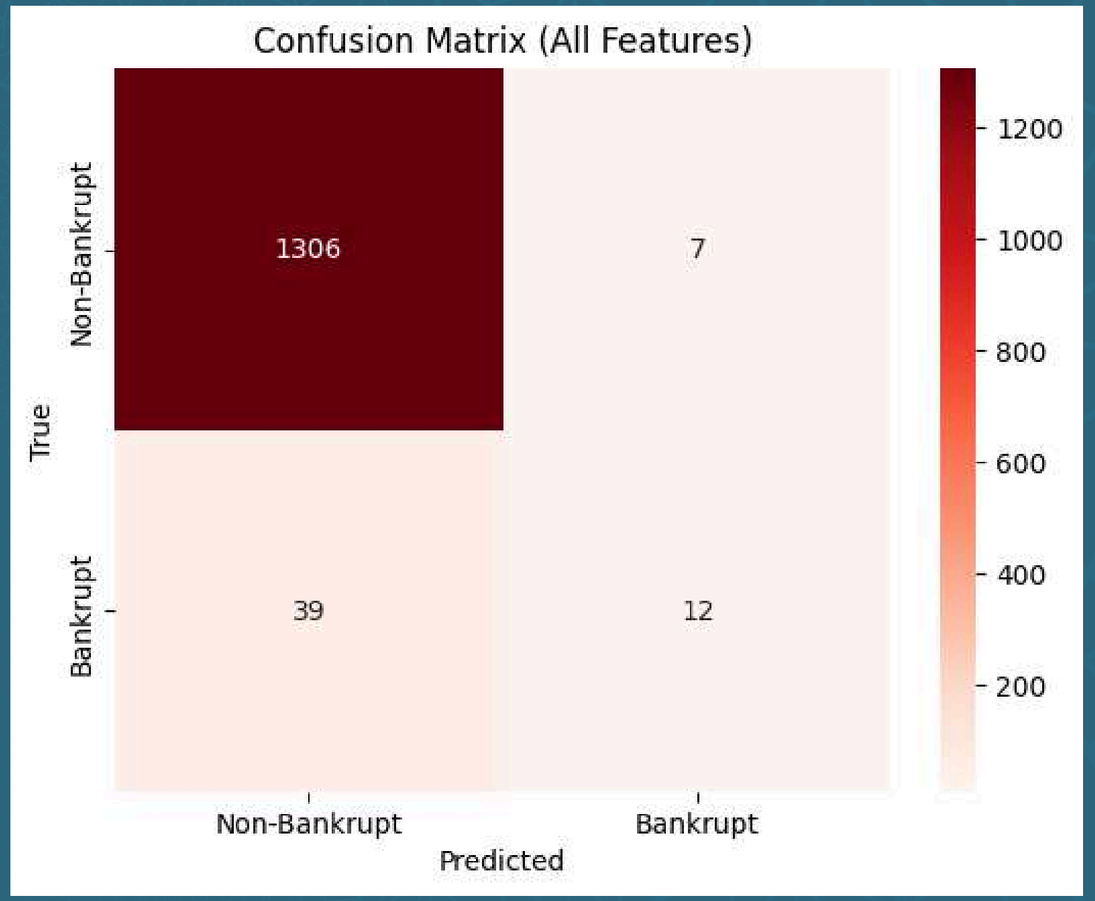
Logistic Regression model with Recursive Feature Elimination (RFE)

LOGISTIC REGRESSION MODEL

- Accuracy: 96.63% with all features.

Confusion Matrix:

- True Negatives (1306): Correctly predicted Non-Bankrupt.
- True Positives (12): Correctly predicted Bankrupt.
- False Negatives (39): Incorrectly predicted Non-Bankrupt when it was Bankrupt.
- False Positives (7): Incorrectly predicted Bankrupt when it was Non-Bankrupt.
- The model performs well in predicting Non-Bankrupt companies (1306 correct predictions).
- It struggles more with detecting Bankrupt companies, with a higher number of False Negatives (39) than False Positives (7).

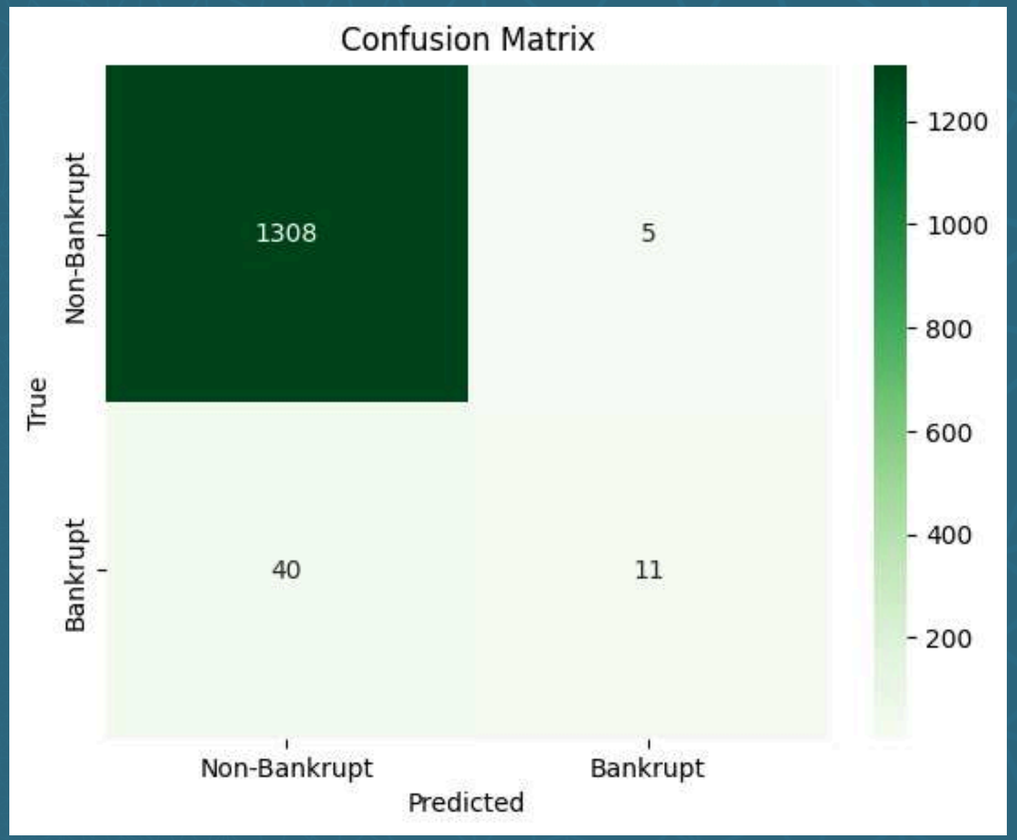


LOGISTIC REGRESSION MODEL

Feature Selection: RFE

After applying feature selection, the confusion matrix shows:

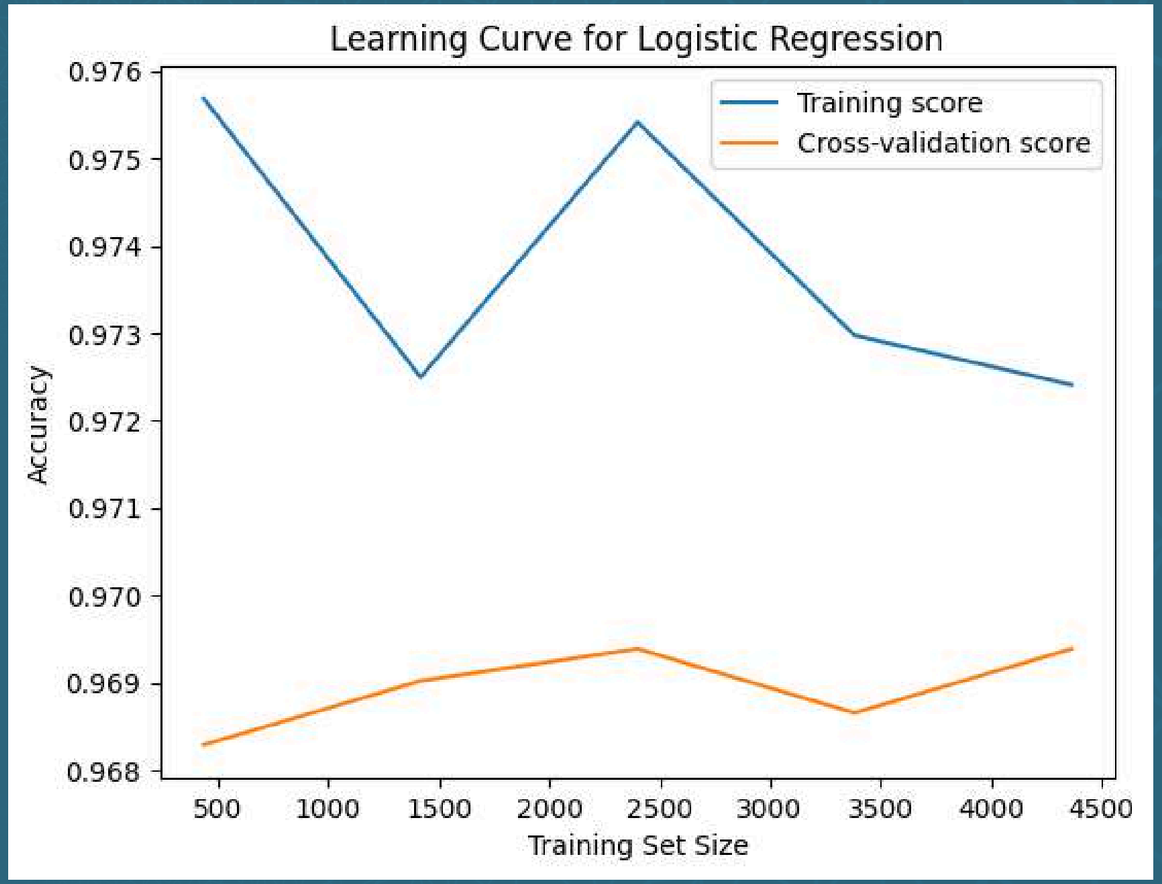
- True Negatives (1308): Correctly predicted Non-Bankrupt.
- True Positives (11): Correctly predicted Bankrupt.
- False Negatives (40): Missed Bankrupt companies (predicted Non-Bankrupt).
- False Positives (5): Incorrectly flagged Non-Bankrupt companies as Bankrupt.
- False Positives decreased, meaning fewer mistakes in predicting Non-Bankrupt as Bankrupt.
- True Positives slightly improved.



LOGISTIC REGRESSION MODEL

. Learning Curve

- The model performs well on the training data.
- The model also does well on new data, both lines get closer



LOGISTIC REGRESSION MODEL

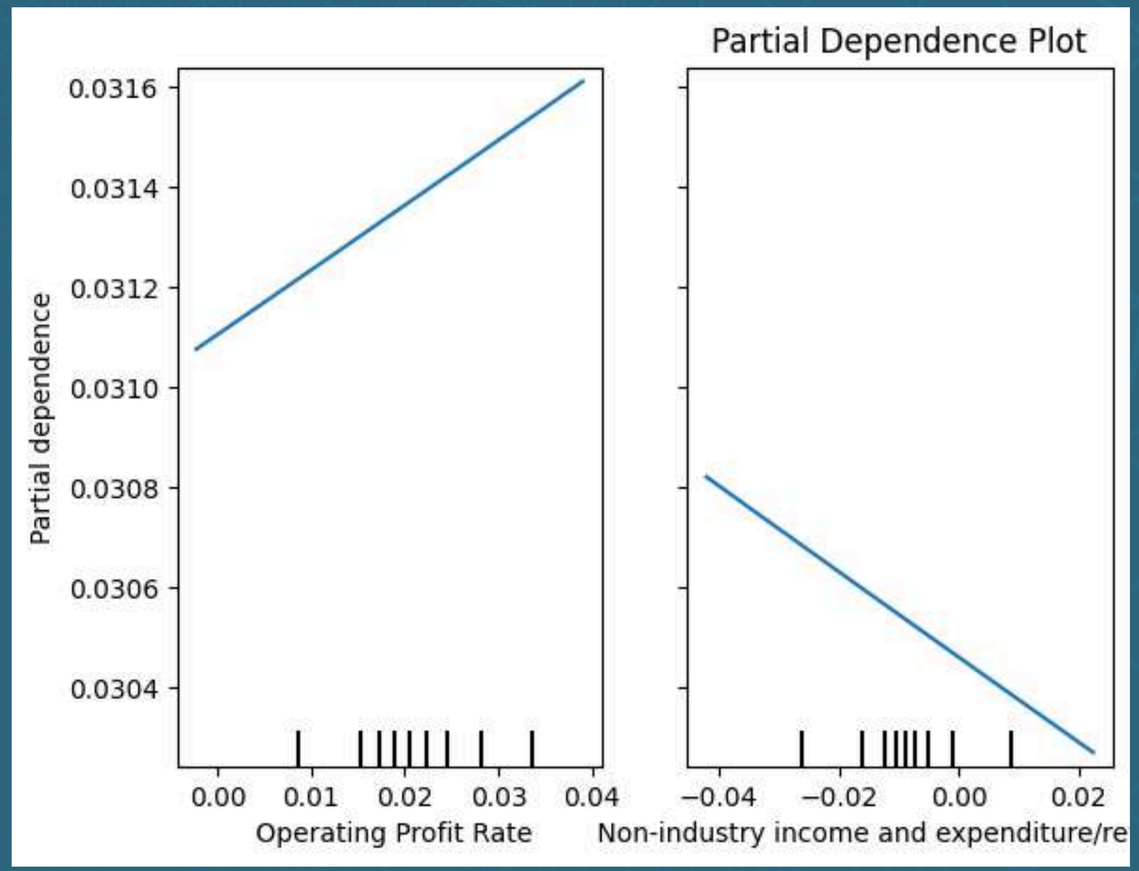
- Interpretation
- Even though there are some negative signs like high debt and negative net worth, the company is still doing well because it has good profits, cash flow, and overall strong financial performance. Because of this, the model predicts with 99% certainty that the company is not bankrupt.

LOGISTIC REGRESSION MODEL

PDP plot

Interpretation

- Operating Profit Rate: As the company's operating profit increases, the chances of it being non-bankrupt go up.
- Non-industry income and expenditure: When the company has higher non-industry income, the chances of it being non-bankrupt decrease, suggesting a higher risk of bankruptcy.



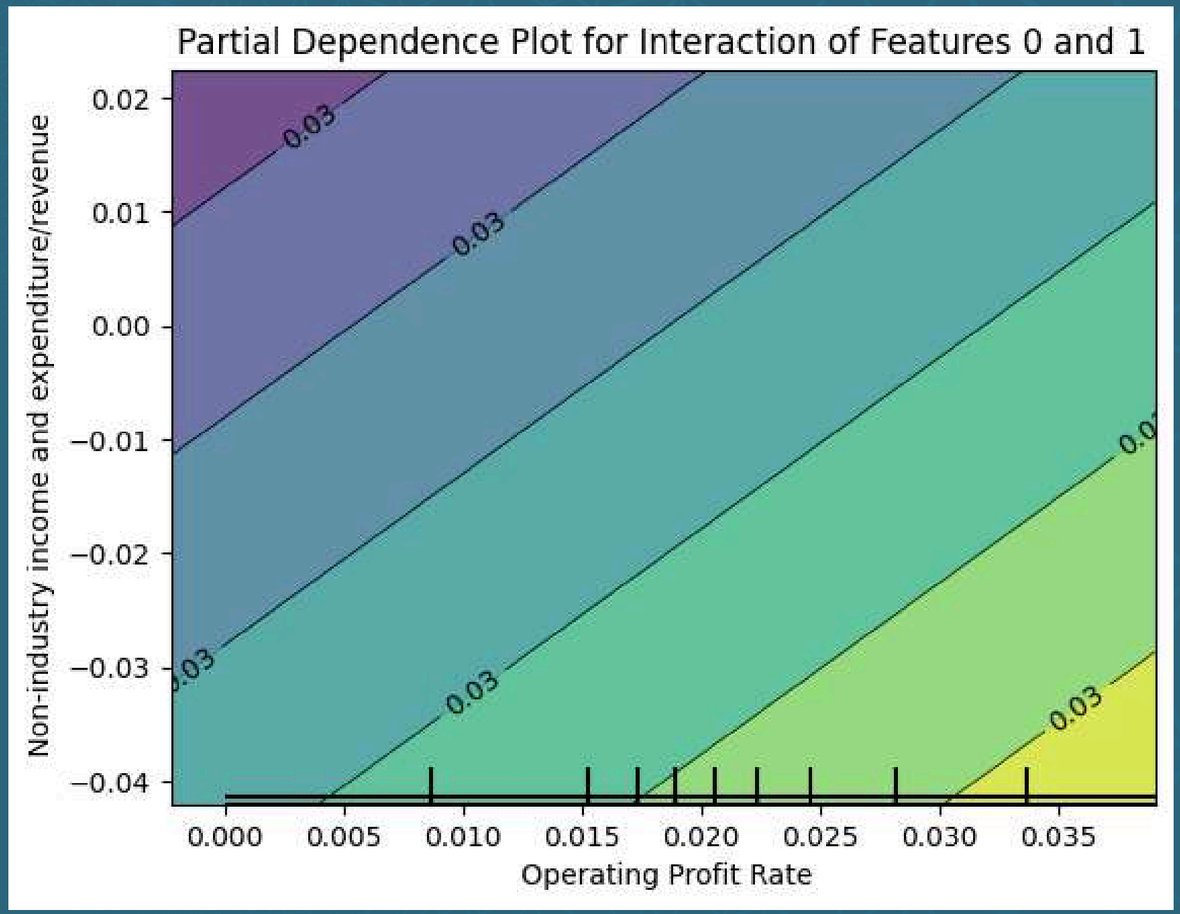
LOGISTIC REGRESSION MODEL

PDP plot

- Plotting Partial Dependence Plot (PDP) for the interaction of features 0 and 1

Interpretation

- The Operating Profit Rate increases, and the Non-industry income also changes, suggesting that the two features influence each other.
- The plot indicates that the combined effect of these two features can either increase or decrease the outcome, depending on their values.

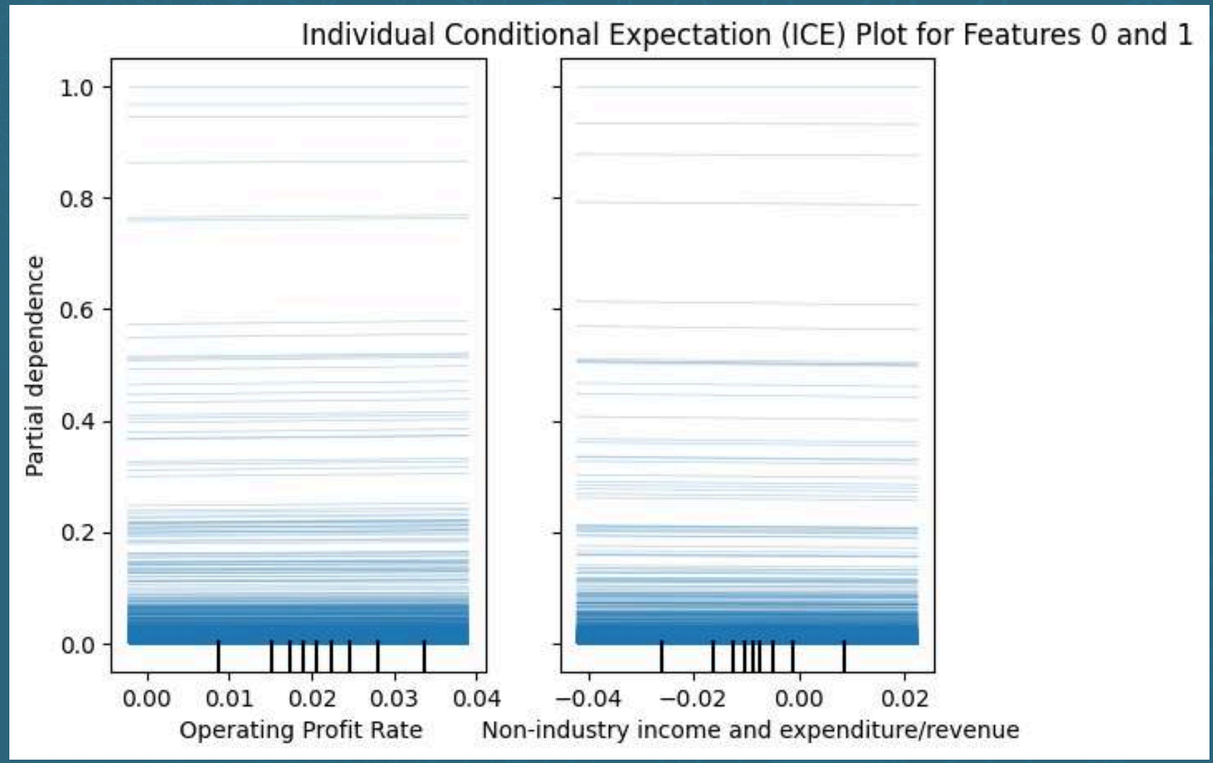


LOGISTIC REGRESSION MODEL

ICE plot

Interpretation

- Operating Profit Rate: As the company's operating profit increases, the chance of it being non-bankrupt usually increases, but this effect can vary for different cases.
- Non-industry income and expenditure: Higher values of non-industry income and expenses tend to lower the chance of being non-bankrupt, but this effect is not the same for every case.



LOGISTIC REGRESSION MODEL

SHAP

LOGISTIC REGRESSION MODEL

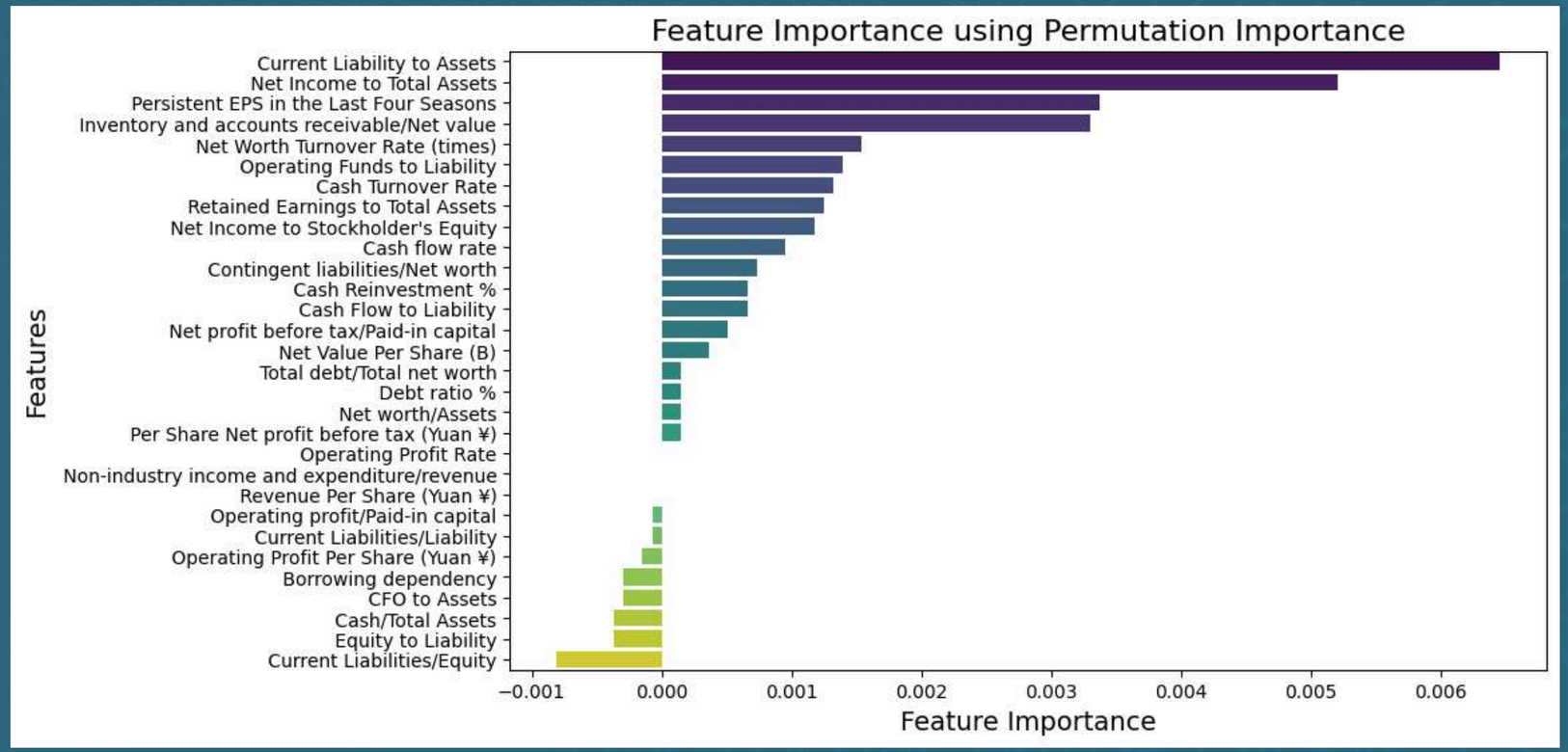
SHAP

- **Interpretation**
- Key Features: Features like Persistent EPS in the Last Four Seasons, Cash/Total Asset, and Net Income to Total Assets have the most significant impact on predictions.
- SHAP Values: Positive SHAP values push predictions towards the positive class (bankruptcy or non-bankruptcy), while negative SHAP values push predictions in the opposite direction.
- Each point on the horizontal axis of the SHAP summary plot represents the effect of a feature across all data points.
- The wider horizontal spread indicates that the feature has a different effect on various data points.
- The narrower horizontal spread shows that the feature has a consistent effect on the data points.
- The vertical spread in the plot shows how the feature values are distributed for the same SHAP value.
- The color of the points indicates the feature value range:
- Red/pink points correspond to higher feature values, which have a stronger effect on the prediction.
- Blue/purple points correspond to lower feature values, which have a weaker effect on the model's prediction.

LOGISTIC REGRESSION MODEL

Permutation importance

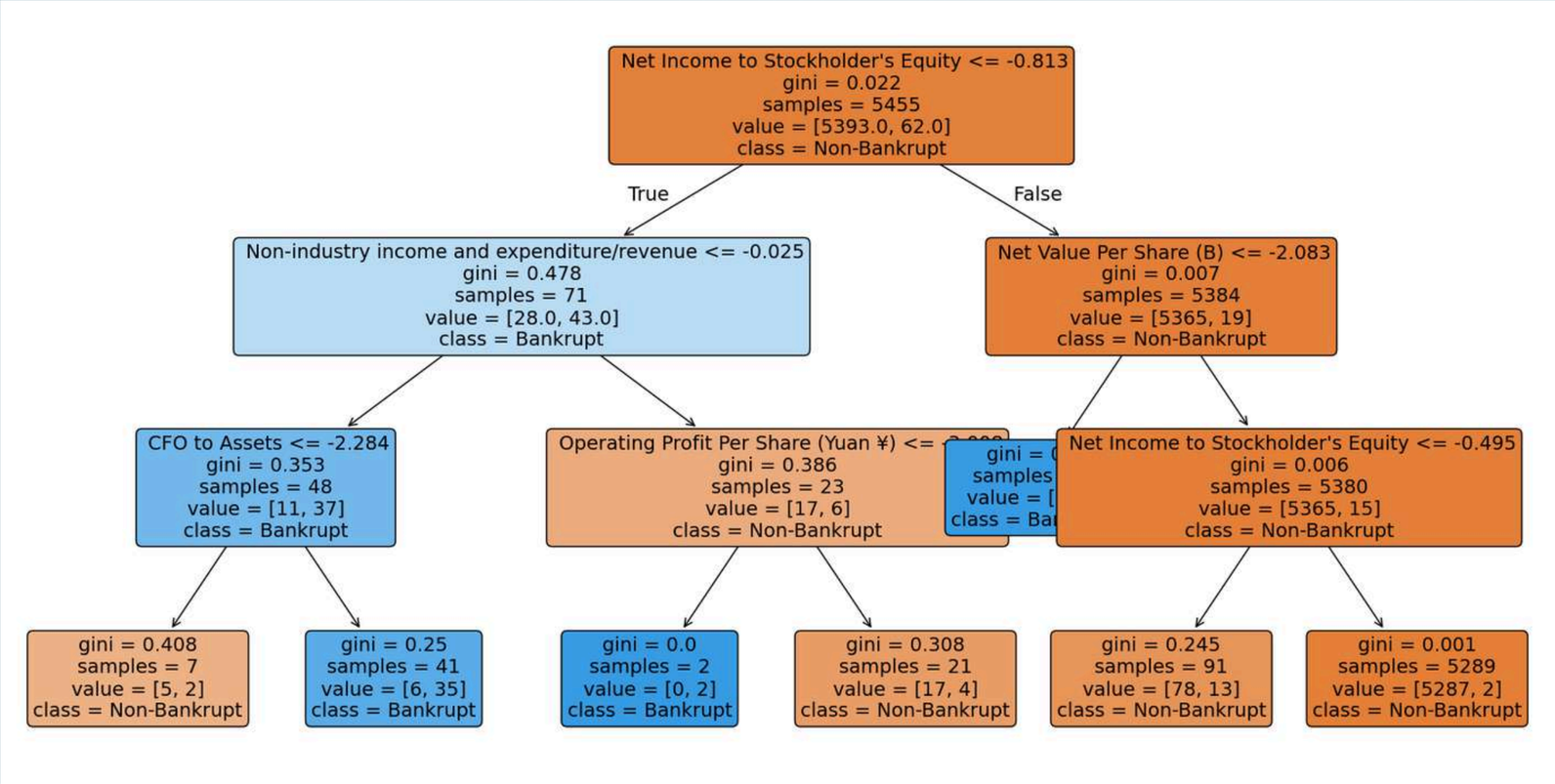
- Features like Current Liability to Assets, Net Income to Total Assets, and Persistent EPS in the Last Four Seasons have the biggest impact on the model's predictions.
- Features like Equity to Liability and Cash/Total Assets have little effect on the predictions and don't add much value to the model.



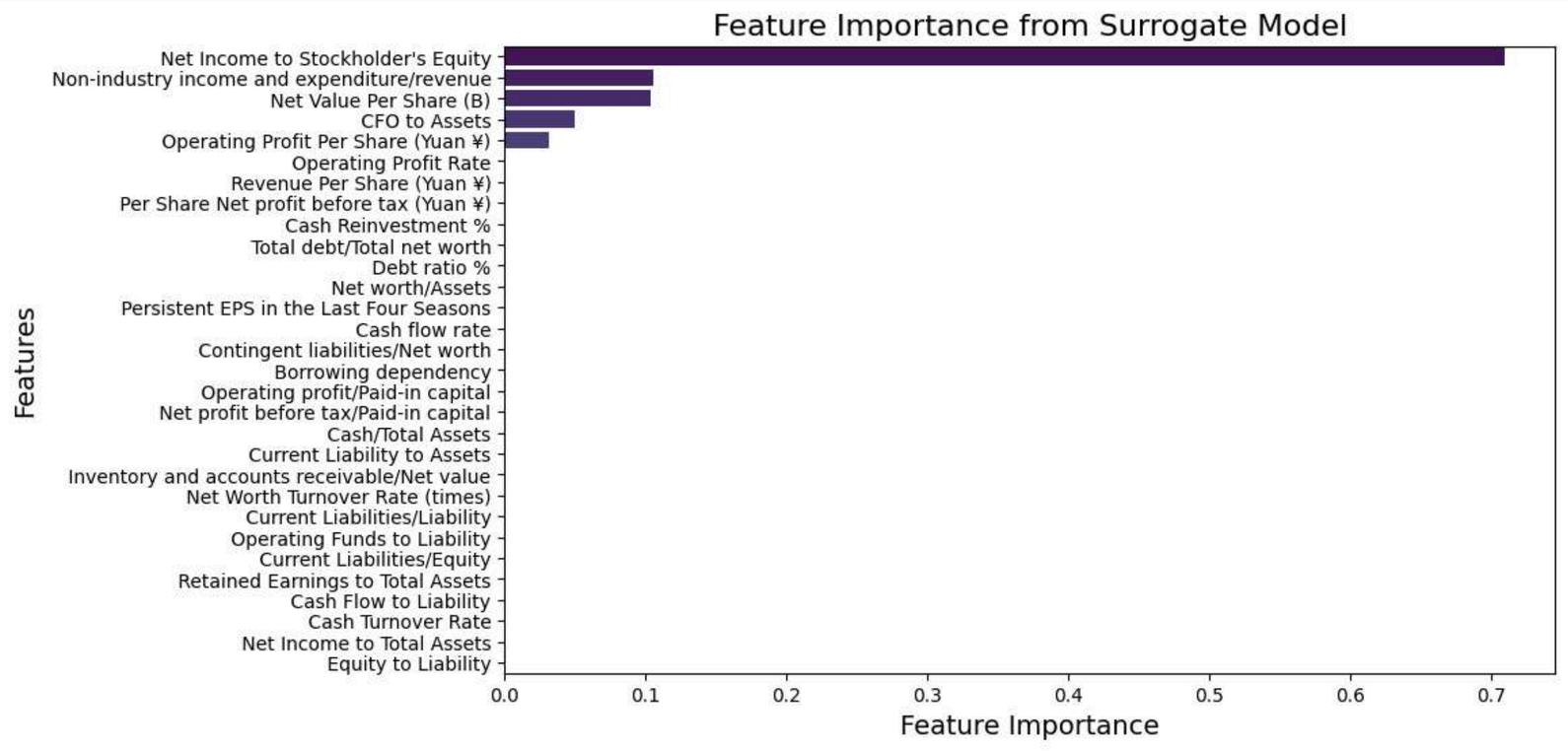
LOGISTIC REGRESSION MODEL

Surrogate Model (Decision Tree)

- Surrogate Decision Tree:
- Purpose: A simpler model that approximates the behavior of the complex Logistic Regression model.
- Accuracy: 99.34% in approximating predictions from Logistic Regression.



- **Interpretation**
- It starts by checking Net Income to Stockholders' Equity.
- Depending on its value, the tree branches into different features like Net Value Per Share and Non-industry income and expenditure/revenue.
- The leaf nodes at the bottom of the tree give the final prediction (Bankrupt or Non-Bankrupt).
- The Gini index shows how pure each decision is.

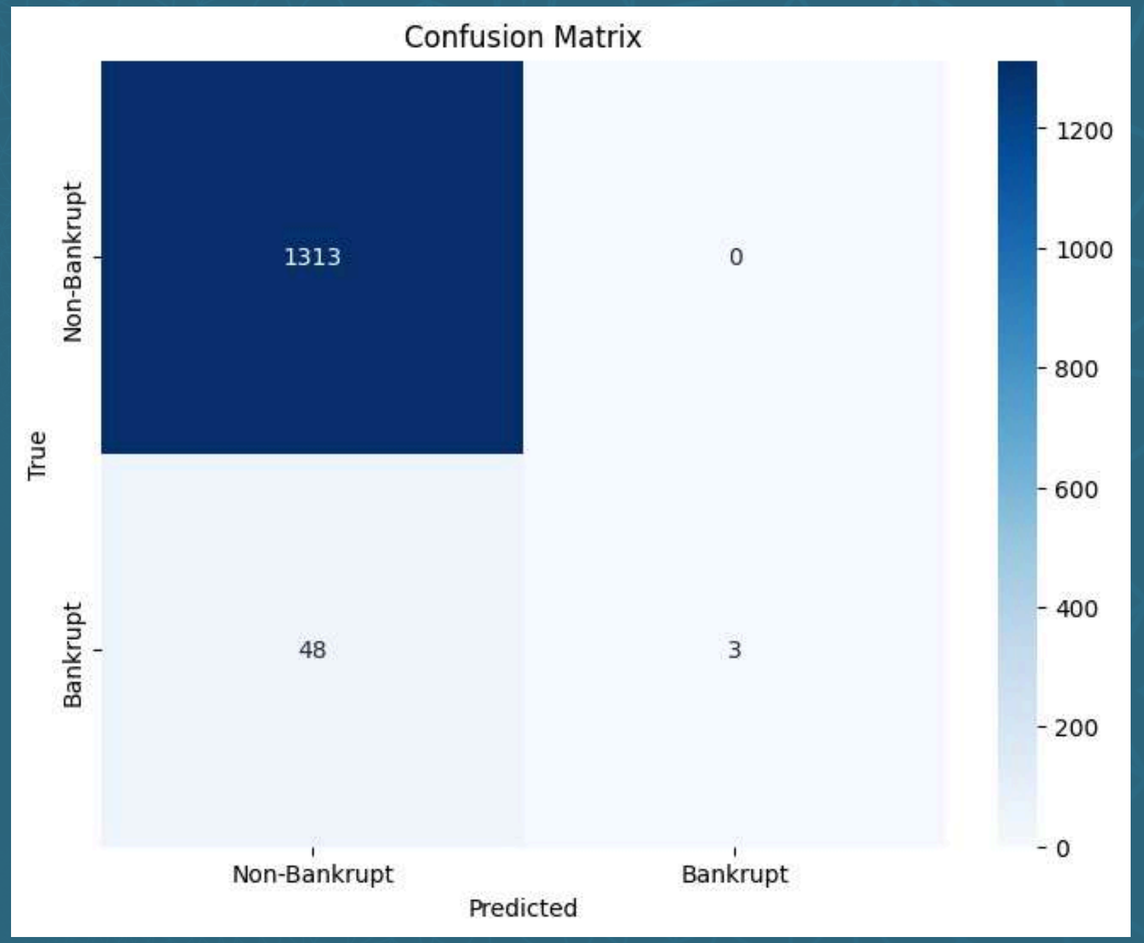


- Based on a simpler model (decision tree) that approximates the original model's behavior.
- Shows how each feature impacts the surrogate model’s predictions.
- The features like "Net Income to Stockholders' Equity" and "Non-industry income and expenditure/revenue" are most important.

ADABOOST MODEL

Accuracy:

- Without feature selection, the accuracy is high (96%), but the recall for predicting bankrupt companies is low (24%).

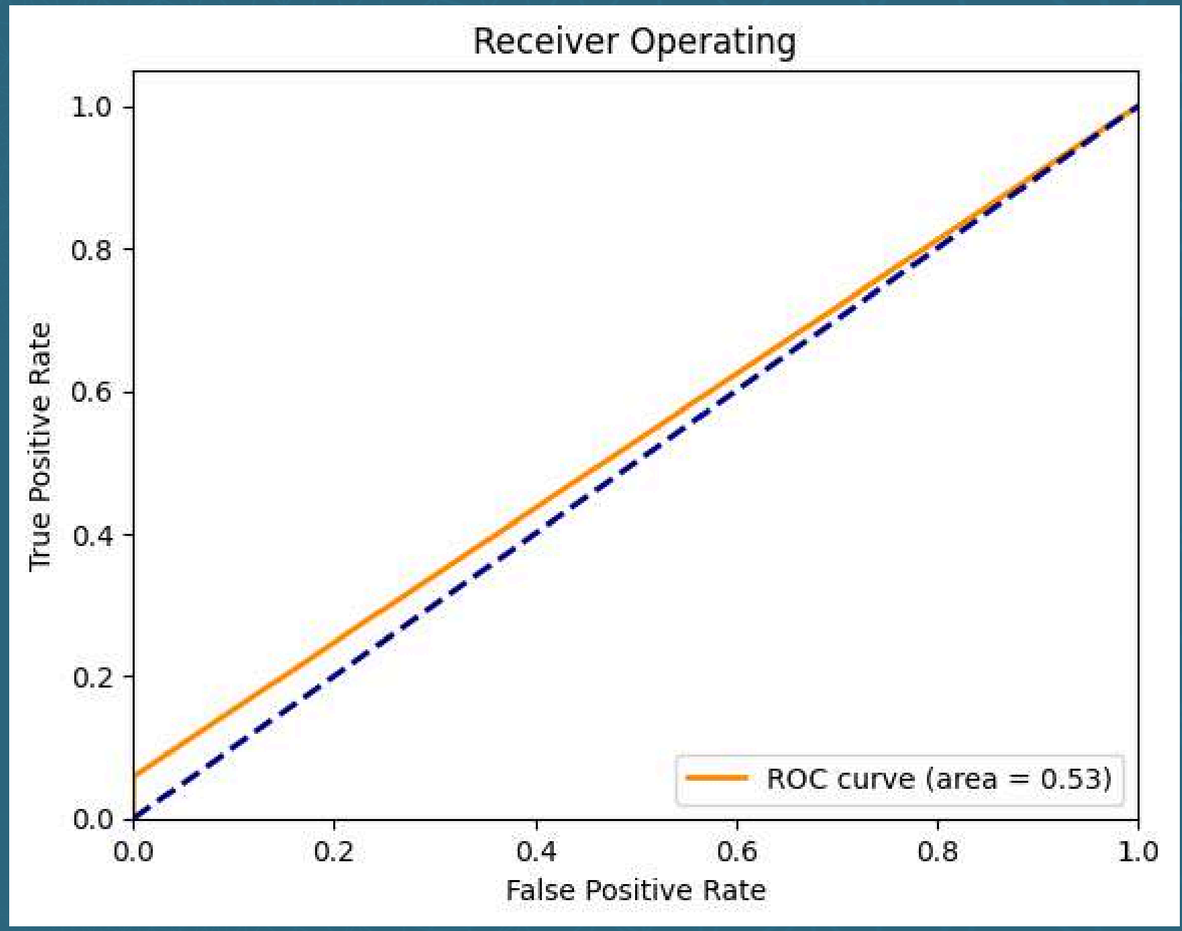


- **Interpretation**
- True Negatives (TN): 1313 correctly predicted non-bankrupt companies.
- False Positives (FP): 0 incorrectly predicted non-bankrupt companies as bankrupt.
- True Positives (TP): 3 correctly predicted bankrupt companies.
- False Negatives (FN): 48 incorrectly predicted bankrupt companies as non-bankrupt.
- The model performs well at predicting non-bankrupt companies (true negatives), but it struggles with bankrupt companies, as it only detected 3 out of 51 bankrupt cases. The false negative rate is high.

ADABOOST MODEL

ROC curve

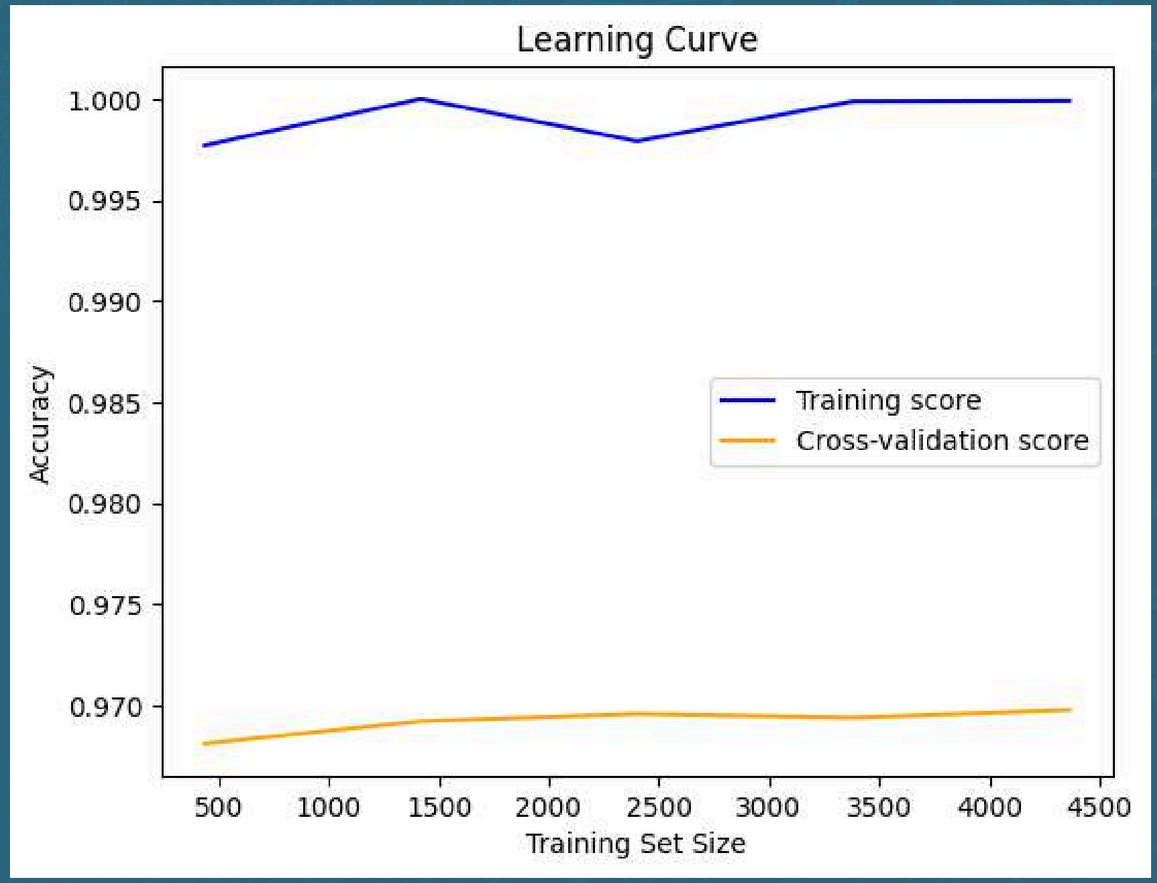
- Interpretation
- Area Under the Curve (AUC): The AUC is 0.53, which suggests that the model performs only slightly better than random guessing. An AUC value close to 0.5 indicates that the model is not distinguishing between the classes effectively.



ADABOOST MODEL

Learning Curve

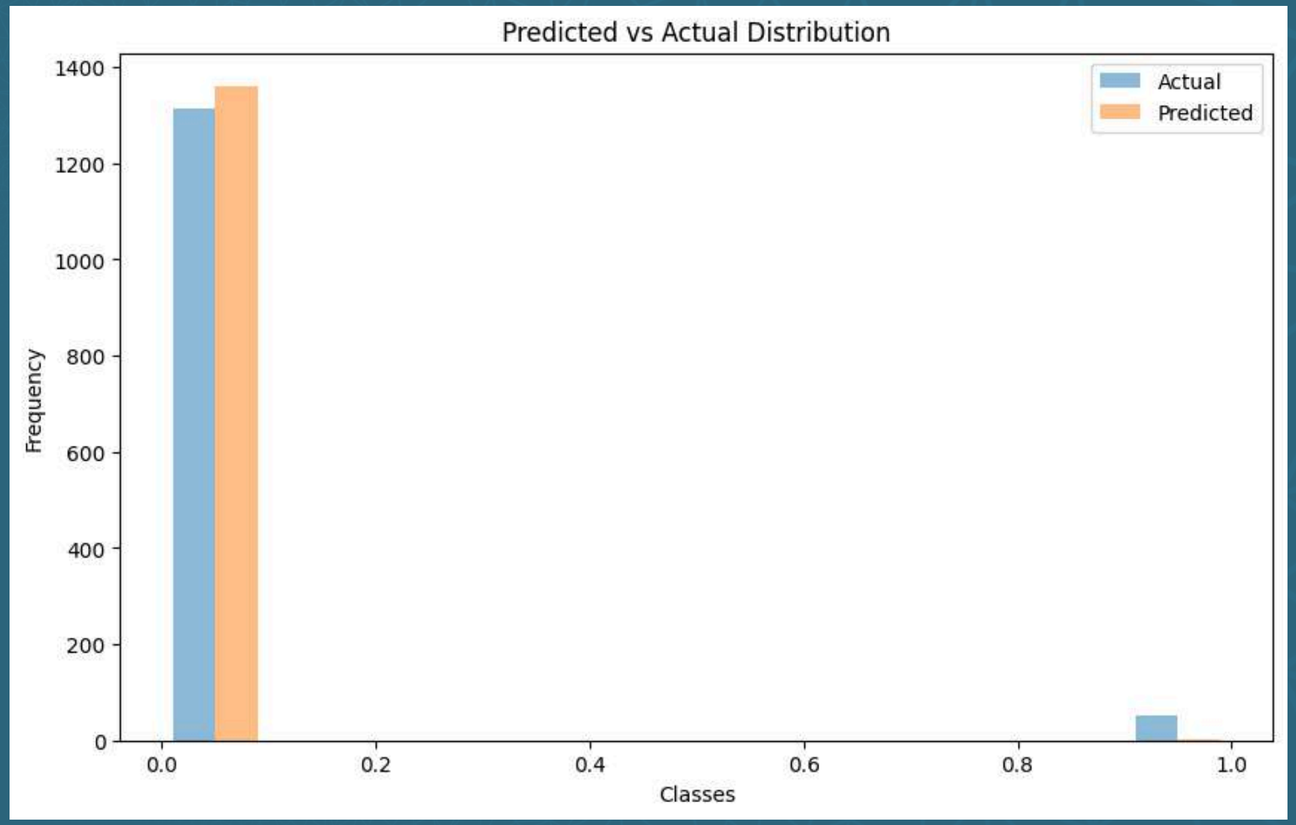
- **Interpretation**
- **Overfitting: The training accuracy is very high, but the cross-validation score is much lower and stable. This suggests that the model may be overfitting to the training data, meaning it is memorizing the training examples rather than generalizing well to unseen data.**



ADABOOST MODEL

Predicted vs Actual Distribution

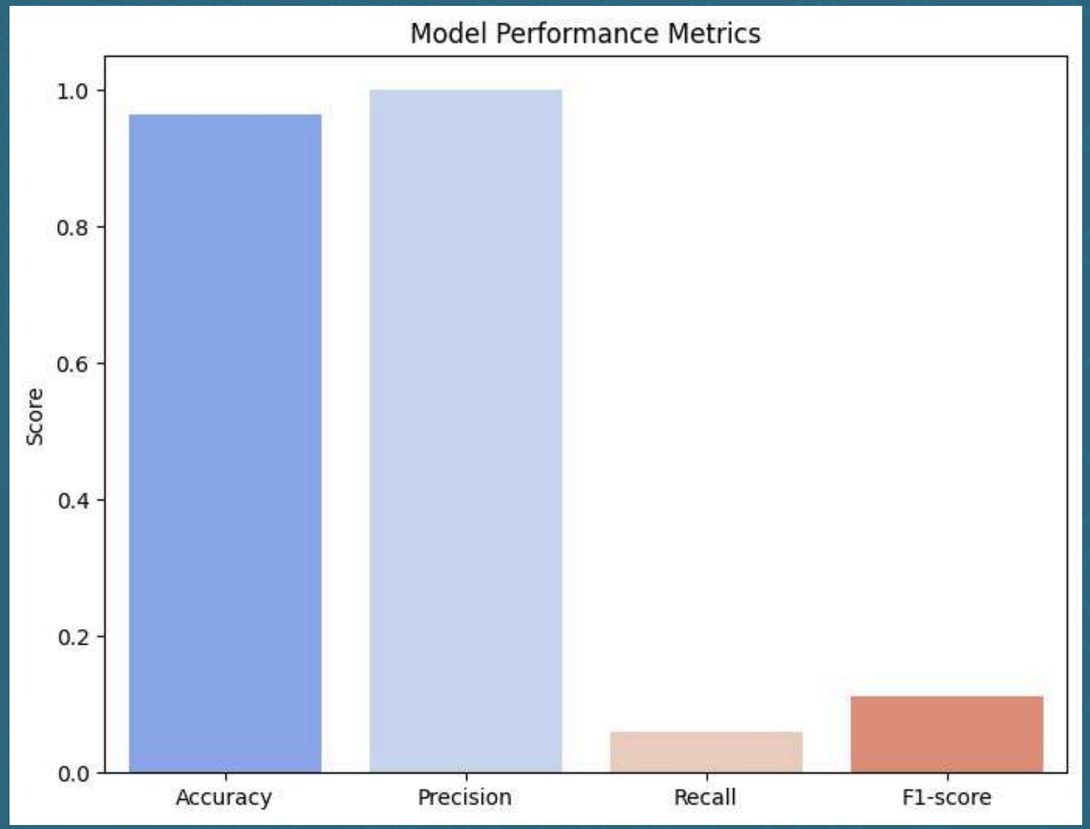
- **Interpretation**
- Overfitting: The training accuracy is very high, but the cross-validation score is much lower and stable. This suggests that the model may be overfitting to the training data, meaning it is memorizing the training examples rather than generalizing well to unseen data.



ADABOOST MODEL

Model Performance Metrics

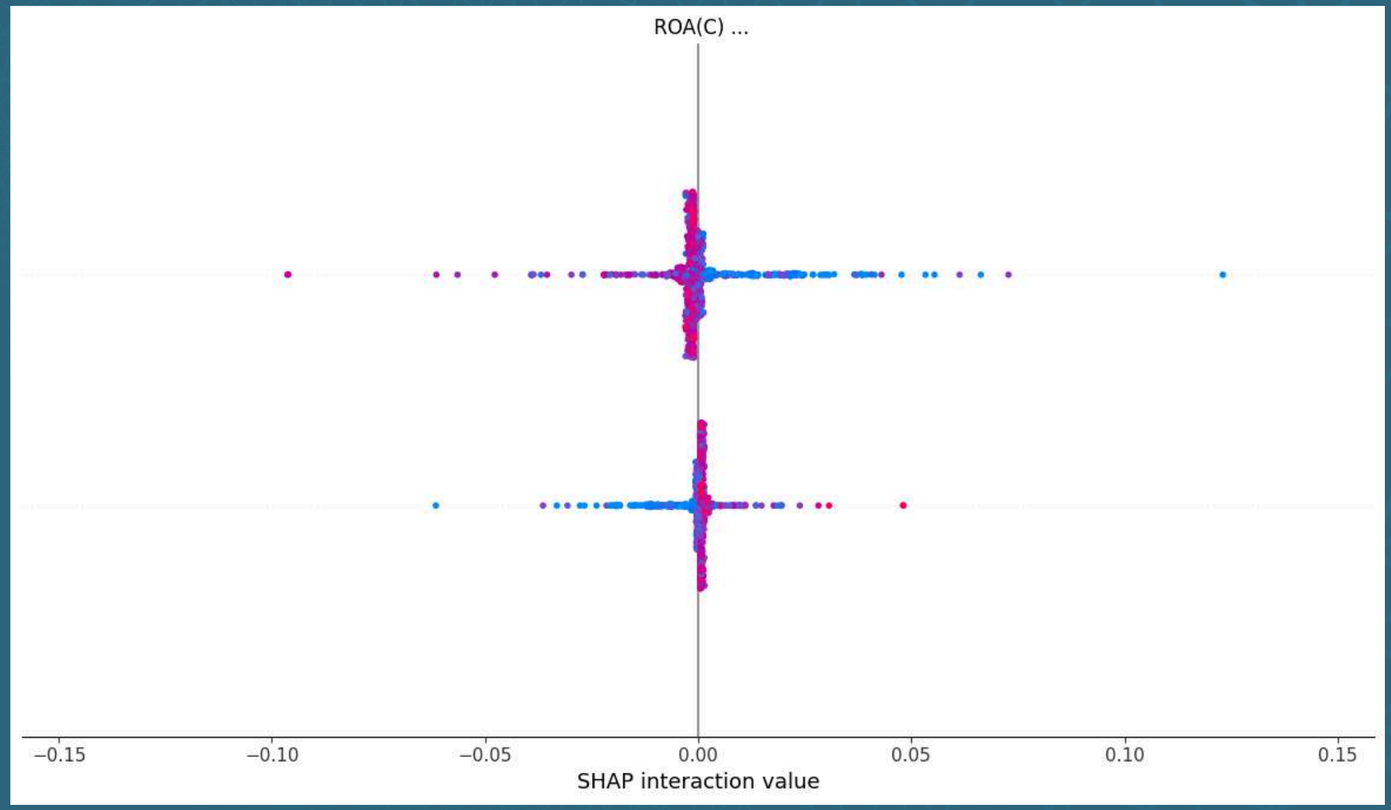
- **Interpretation**
- High Accuracy is not indicative of good performance because the model is likely predicting the class (Non-Bankrupt) most of the time.
- Poor Recall and F1-score show that the model is struggling to detect bankrupt companies.



ADABOOST MODEL

Predicted vs Actual Distribution

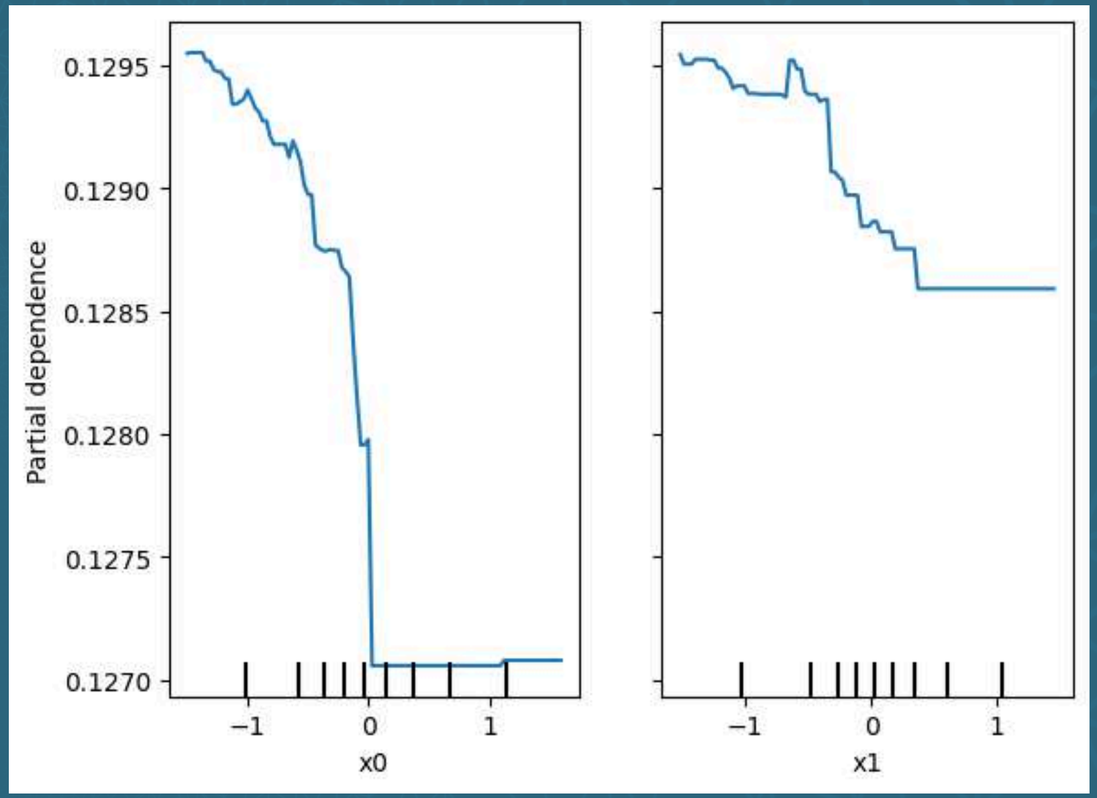
- **Interpretation**
- SHAP interaction value: This plot shows how the interaction between the feature (ROA(C)) and the model's prediction contributes to the final output.
- X-axis: Represents the SHAP interaction values.
- Y-axis: The specific feature (ROA(C)) that interacts with other features.
- Color of points: the color indicates the value of the feature, with blue indicating lower values and pink indicating higher values of ROA(C).
- Spread of Data Points: Most of the data points are around the center so the feature ROA(C) has a limited effect on the model's output for most instances.
- Feature Importance: The SHAP interaction values suggest that ROA(C) is interacting with other features in a way that influences the model's prediction.



ADABOOST MODEL

PDP Plot

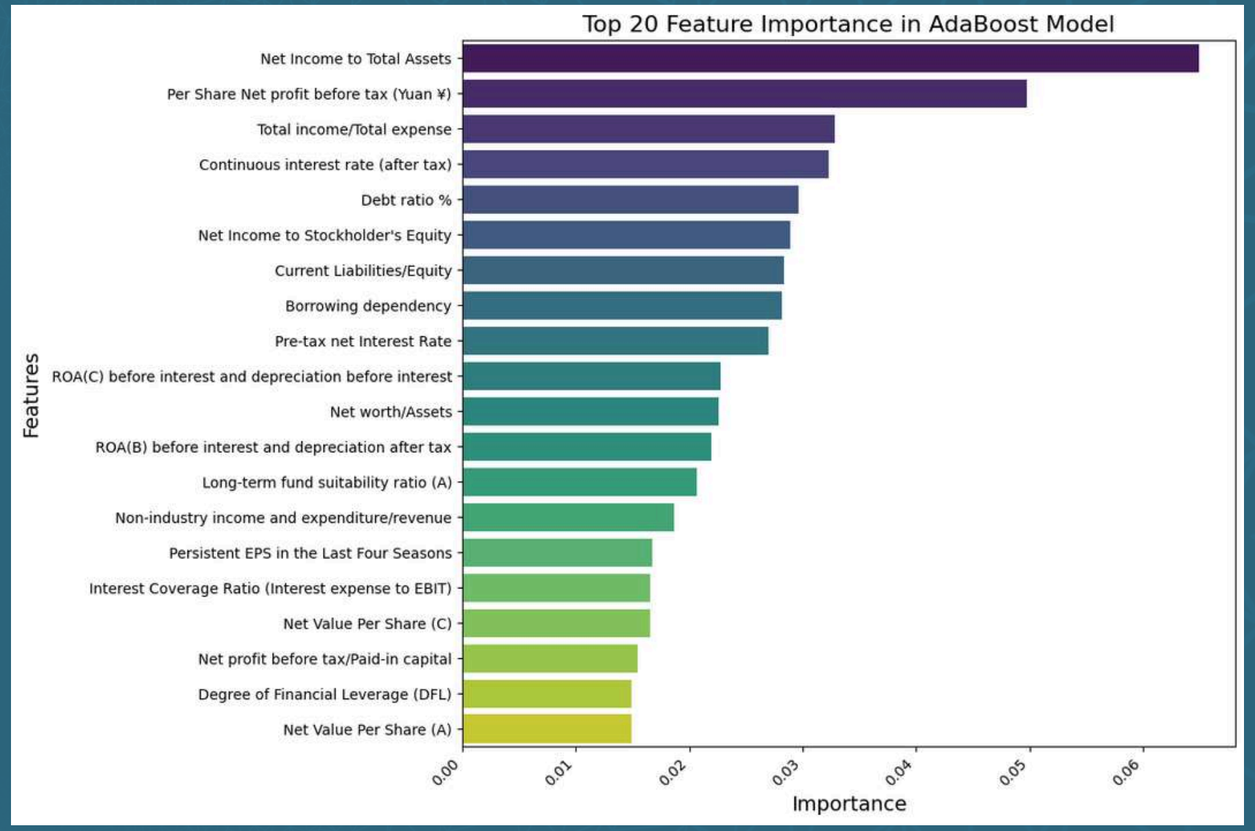
- **Interpretation**
- Feature x0: When x0 is near 0, small changes cause big changes in the model’s prediction.
- Feature x1: As x1 changes, the model’s prediction changes smoothly. This means that x1 affects the prediction in a more predictable and continuous .



ADABOOST MODEL

PDP Plot

- **Interpretation**
- Most Important Features:
- The feature Net Income to Total Assets has the highest importance, meaning it plays the biggest role in predicting the model's output.
- The second most important feature is Per Share Net profit before tax (Yuan ¥), followed by **Total income/Total expense, which also significantly influences the predictions.

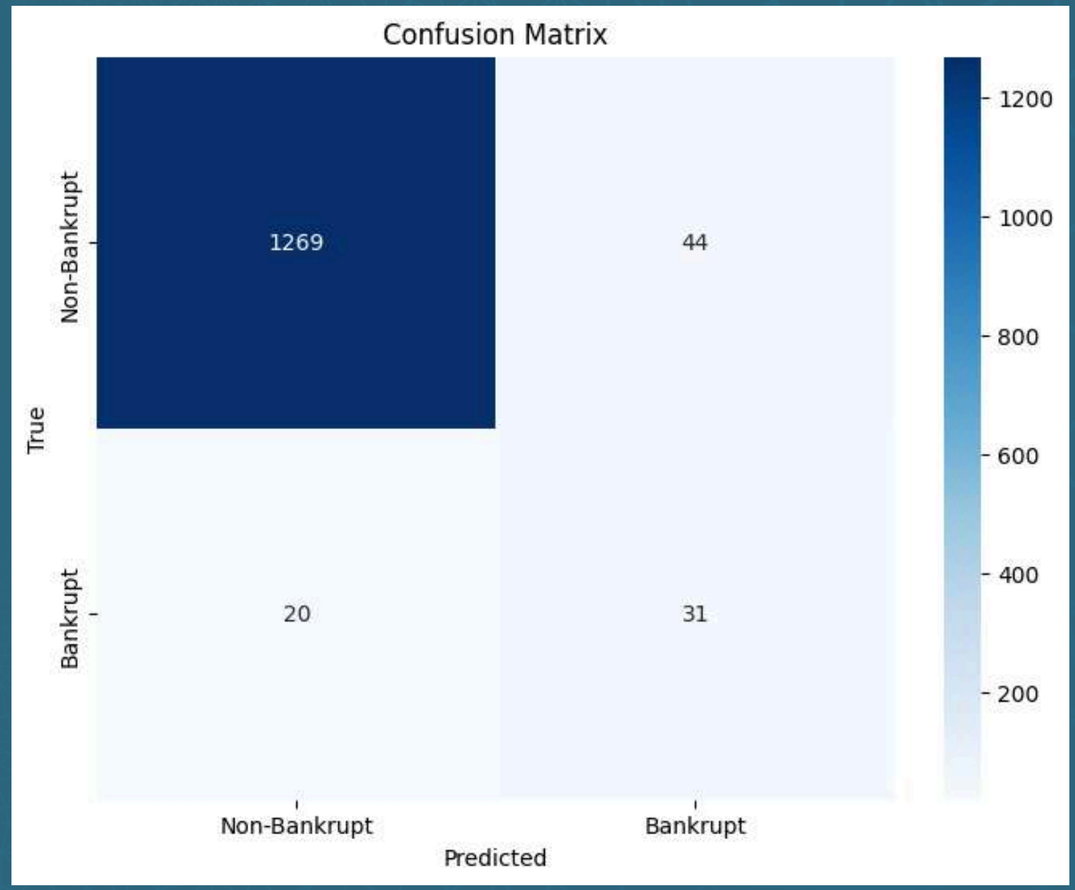


ADABOOST MODEL

SMOTE for Balancing Classes

- The application of SMOTE helped balance the dataset by generating synthetic examples of bankrupt companies.

- **The model's performance improved:**
- Precision: 41.33%
- Recall: 60.78% (from 5.88%)
- F1-Score: 49.21% (improvement).

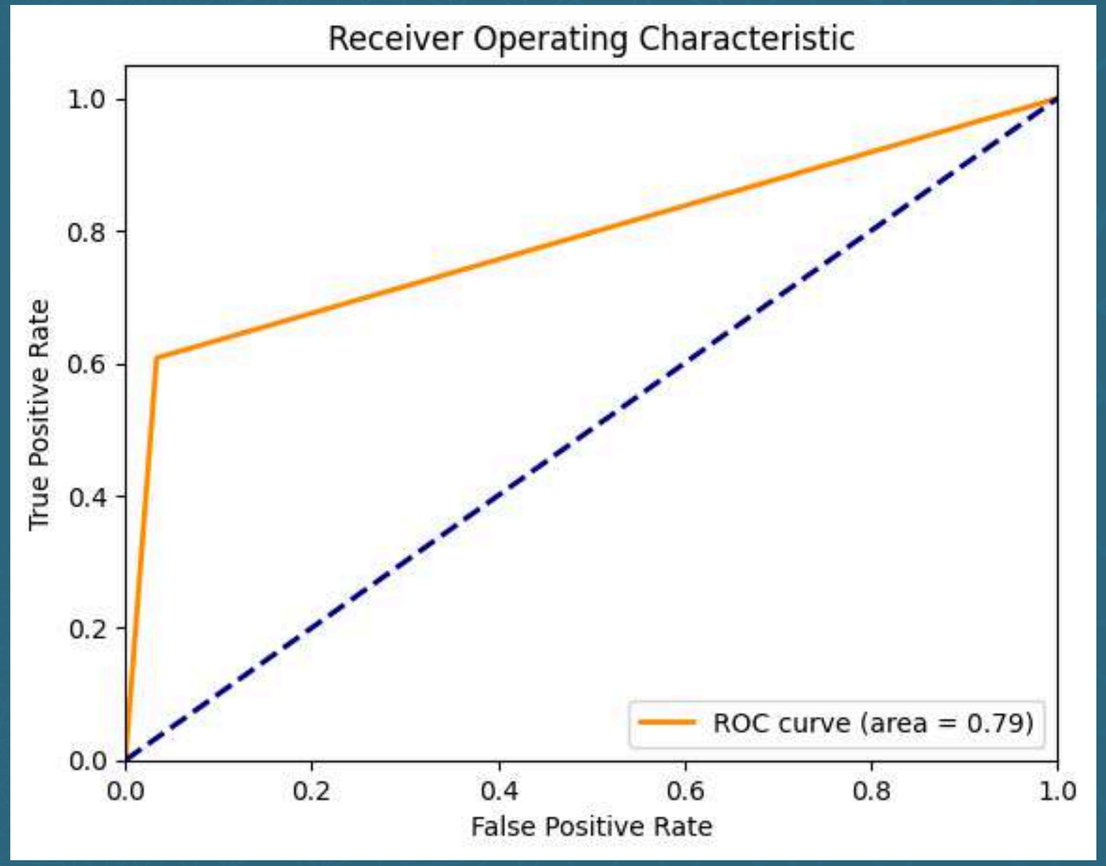


ADABOOST MODEL

SMOTE for Balancing Classes

- The application of SMOTE helped balance the dataset by generating synthetic examples of bankrupt companies.

- The ROC curve now has a significantly higher AUC of 0.79, which means the model's ability to discriminate between bankrupt and non-bankrupt companies is much better.
- The ROC curve improvement reflects better model performance in both detecting bankrupt companies and avoiding false positives.



MACHINE LEARNING MODELS

GRADIENT BOOSTING (GB): CREATES AN ENSEMBLE OF MODELS SEQUENTIALLY, WHERE EACH MODEL TRIES TO CORRECT THE ERRORS MADE BY THE PREVIOUS ONE, FOCUSING ON REDUCING THE OVERALL ERROR.

BAGGING (BAG): IMPROVES MODEL ACCURACY BY TRAINING MULTIPLE INSTANCES OF THE SAME MODEL ON DIFFERENT RANDOM SUBSETS OF THE TRAINING DATA AND AVERAGING THEIR PREDICTIONS TO REDUCE VARIANCE.

RANDOM FOREST :BUILDS MULTIPLE DECISION TREES AND COMBINES THEIR PREDICTIONS TO IMPROVE ACCURACY AND REDUCE OVERFITTING. IT USES RANDOM SUBSETS OF FEATURES AND DATA TO CREATE DIVERSE TREES, MAKING IT MORE ROBUST AND ACCURATE.

GRADIENT BOOSTING (GB)

```
Accuracy: 0.967741935483871
Precision: 0.6666666666666666
Recall: 0.27450980392156865
```

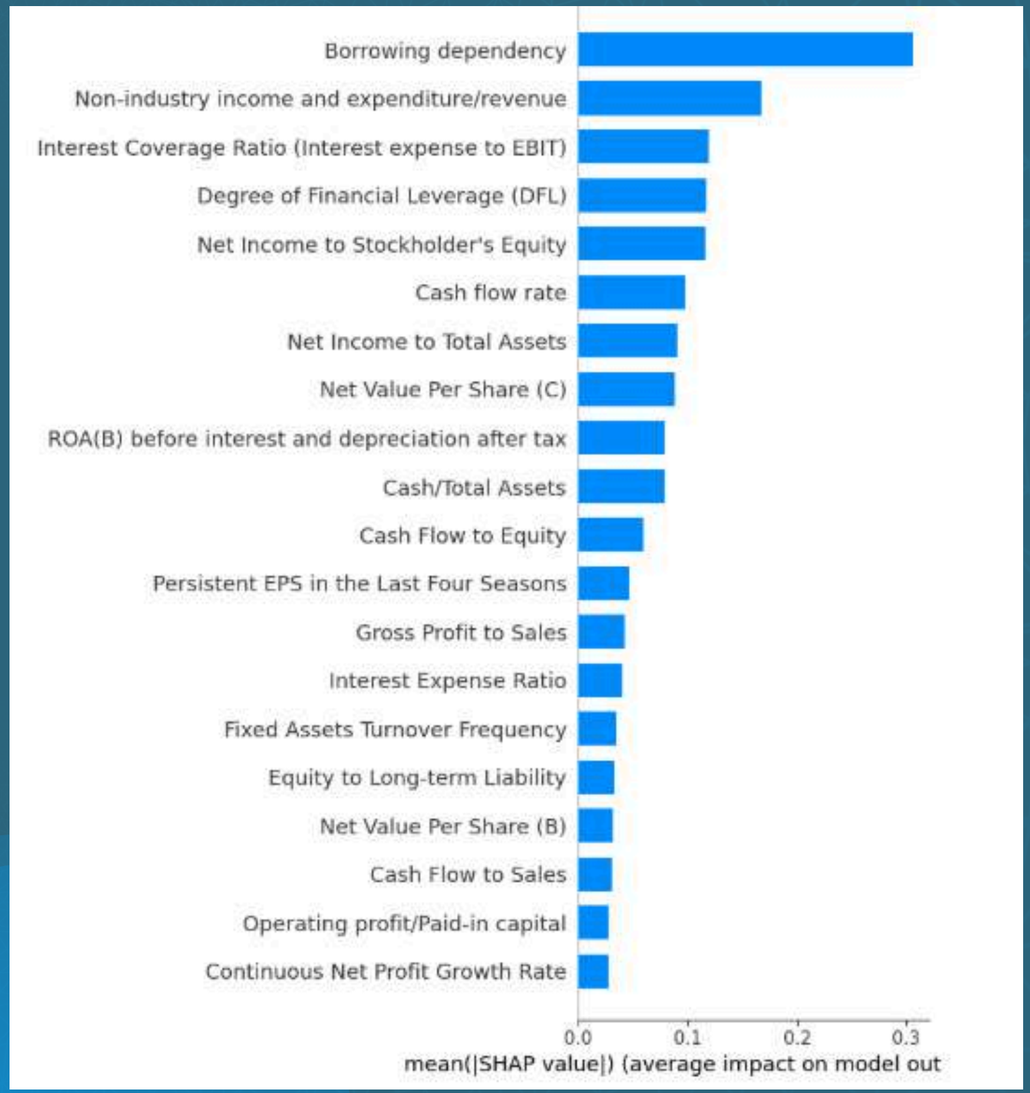
- Accuracy: 0.96 – The model makes correct predictions on most of the data.
- Precision: 0.66 – When the model predicts class 1, it's correct 66% of the time, showing moderate reliability.
- Recall: 0.27 – The model only detects 27% of actual class 1 cases, indicating it misses most positives.

EXPLAINABILITY TECHNIQUES GRADIENT BOOSTING (GB)

1- Feature Importance
2- LIME
3- PDP
4- ICE

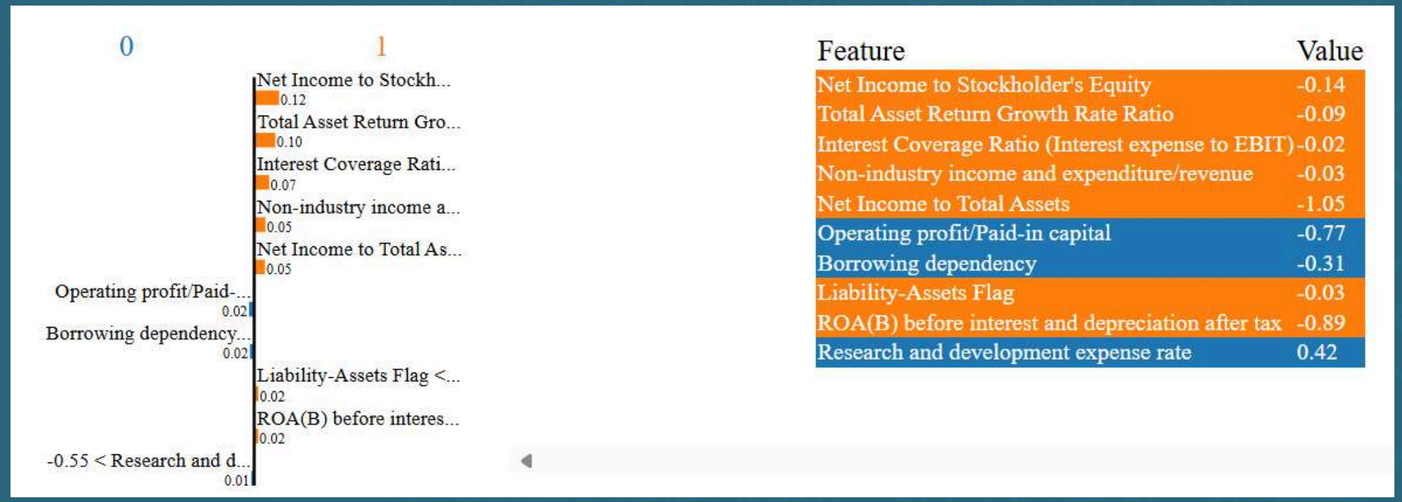
I- FEATURE IMPORTANCE

The feature importance plot displays the most greater top 30 features from the data that affect the performance of the model and have a powerful impact on the training and validation



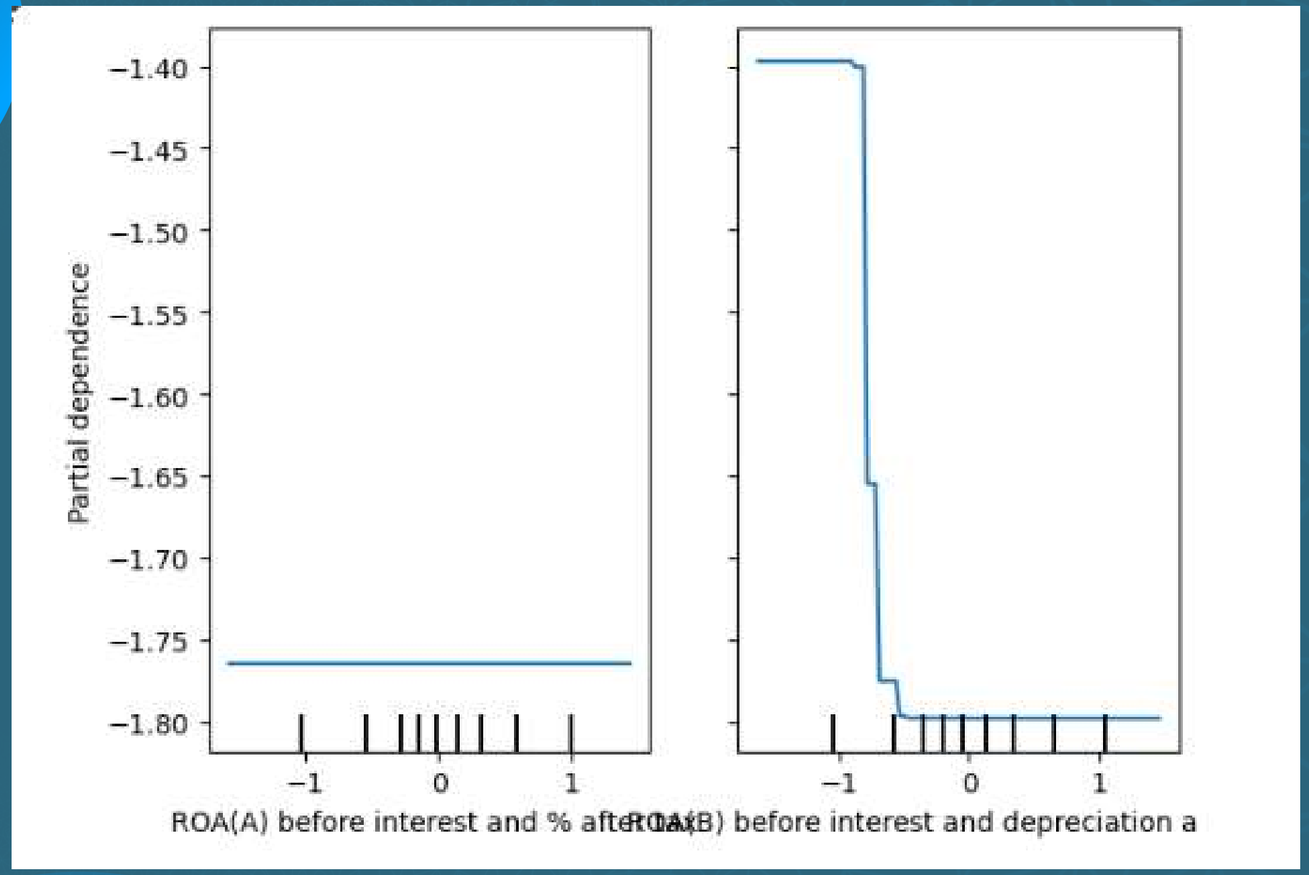
2- LIME

LIME explains individual predictions The output shows which features contributed most to the prediction for this specific instance and whether their contribution pushed the prediction towards 'Bankrupt' or 'Not Bankrupt'.



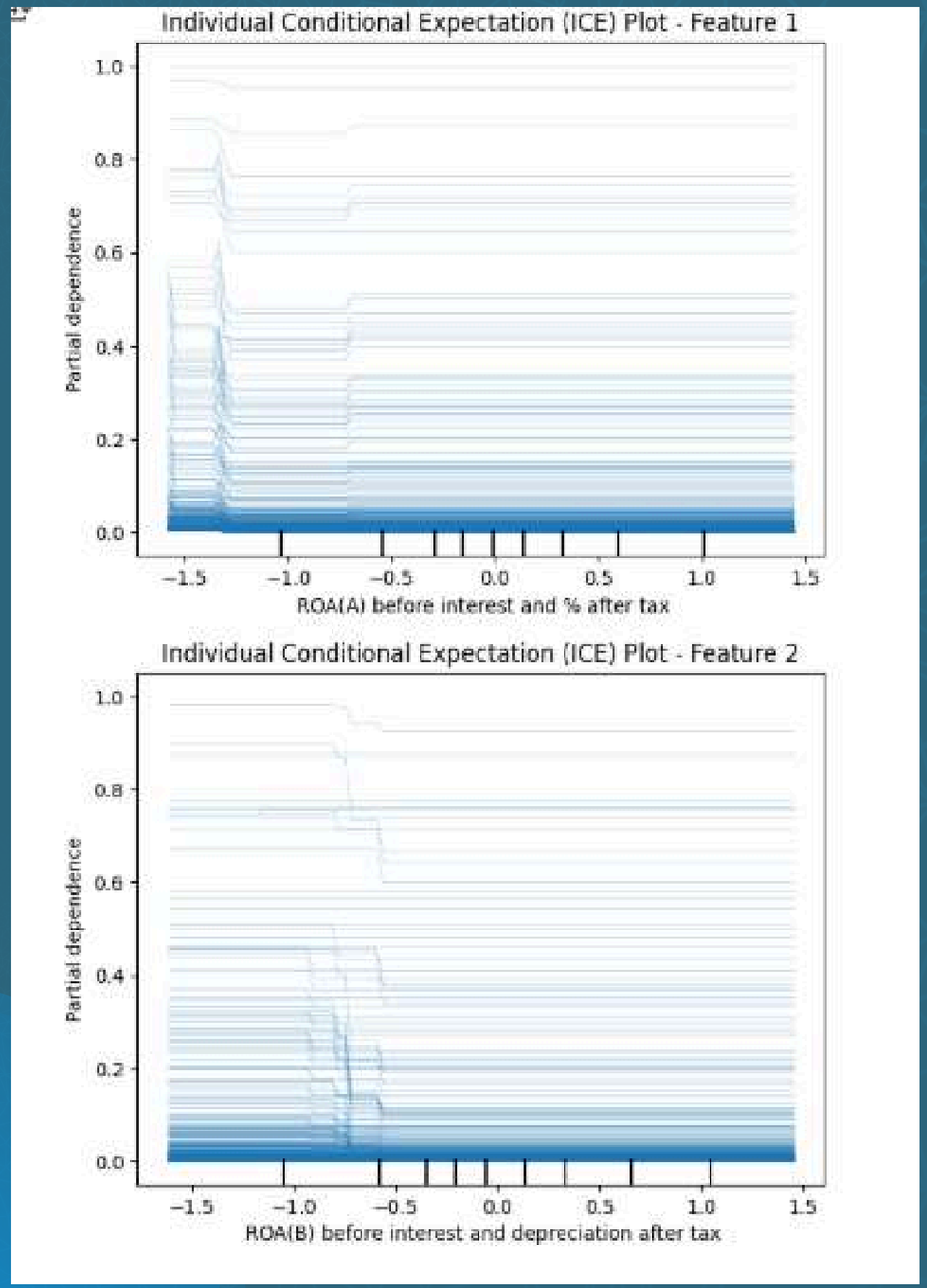
3- PDP

PDP for two features, illustrating how each feature affects the model's prediction. Sharp changes in the plots indicate significant influence of those features on the predicted outcome.



4- ICE

ICE plots show how a model’s prediction changes for an individual data instance as a single feature varies, keeping all other features fixed. provides insight into how different individuals are affected by a feature for each value, a line in the plot.



BAGGING (BAG):

```
Bagging Classifier Accuracy: 0.9618768328445748
Bagging Classifier Precision: 0.4545454545454543
Bagging Classifier Recall: 0.09803921568627451
```

- Accuracy: 0.96 – The model makes correct predictions on most of the data.
- Precision: 0.45– When the model predicts class 1, it's correct 45% of the time, showing moderate reliability.
- Recall: 0.98 – The model only detects 27% of actual class 1 cases, indicating it misses most positives.

EXPLAINABILITY TECHNIQUES BAGGING (BAG):

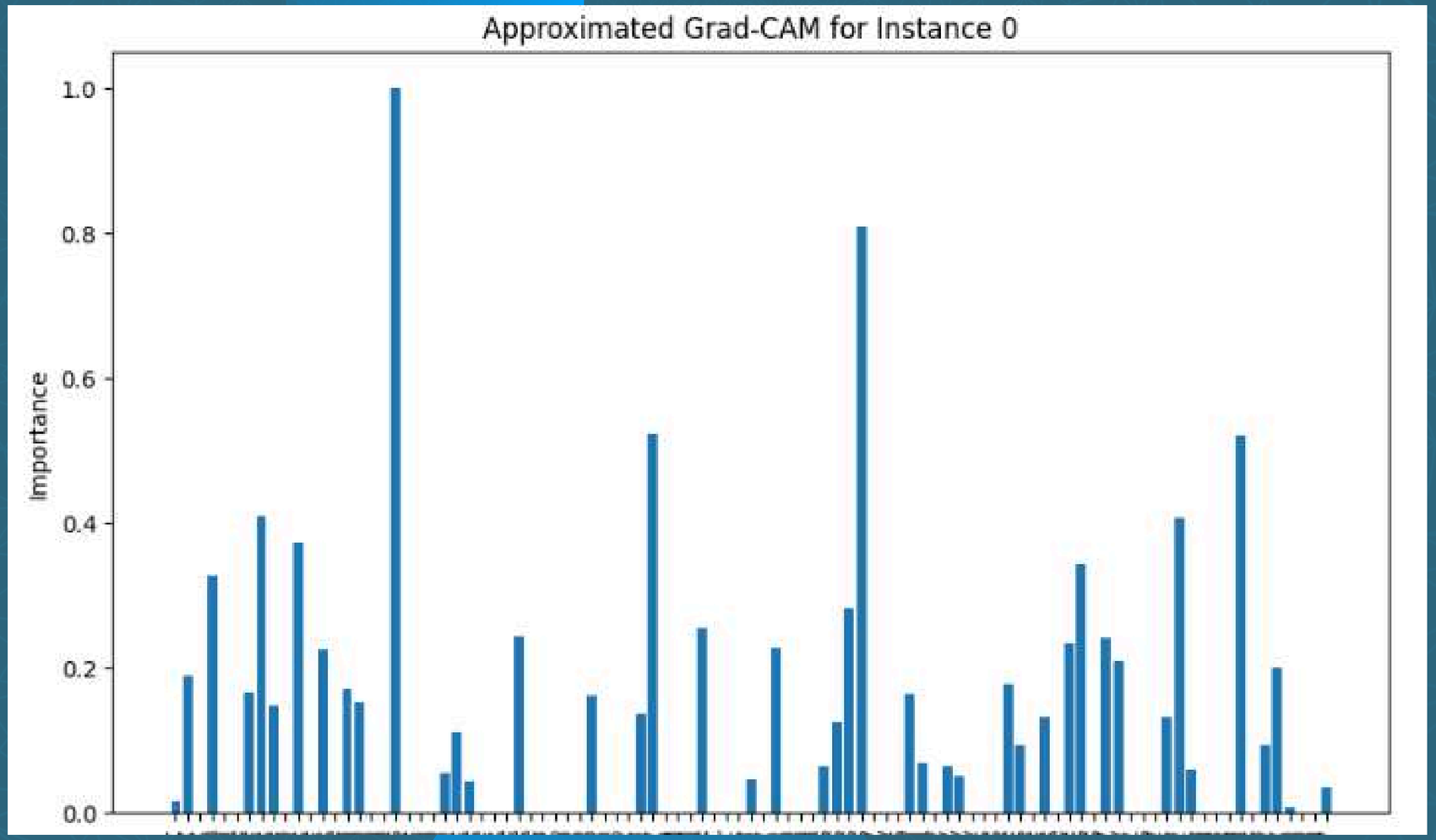
1- Grad-Cam

2-Feature Importance

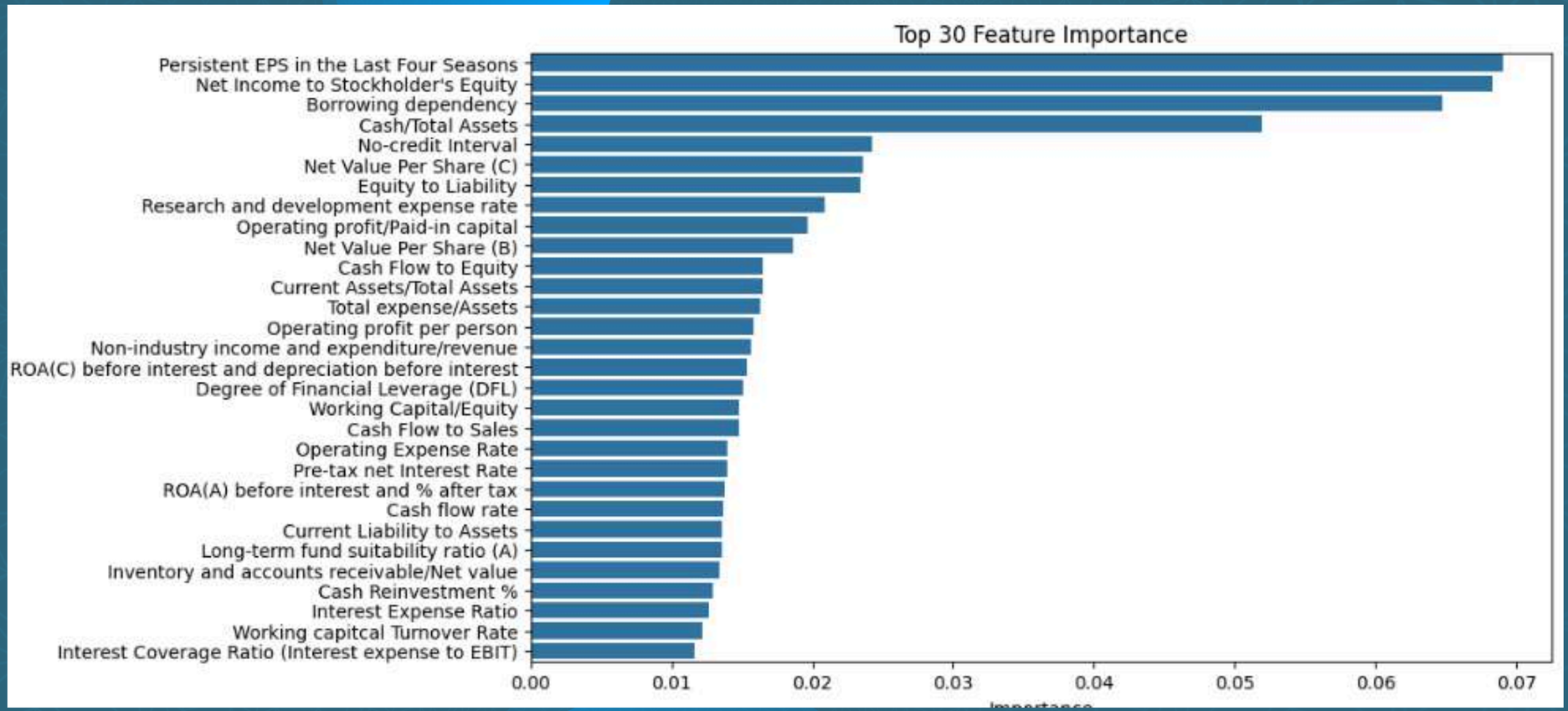
3- LIME

4- permutation_importance

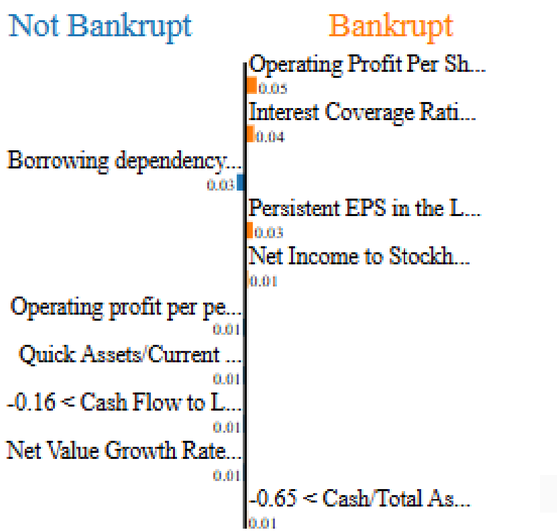
I- GRAD-CAM


Approximated Grad-CAM for Instance 0
Importance
1.0
0.8
0.6
0.4
0.2
0.0

2-FEATURE IMPORTANCE

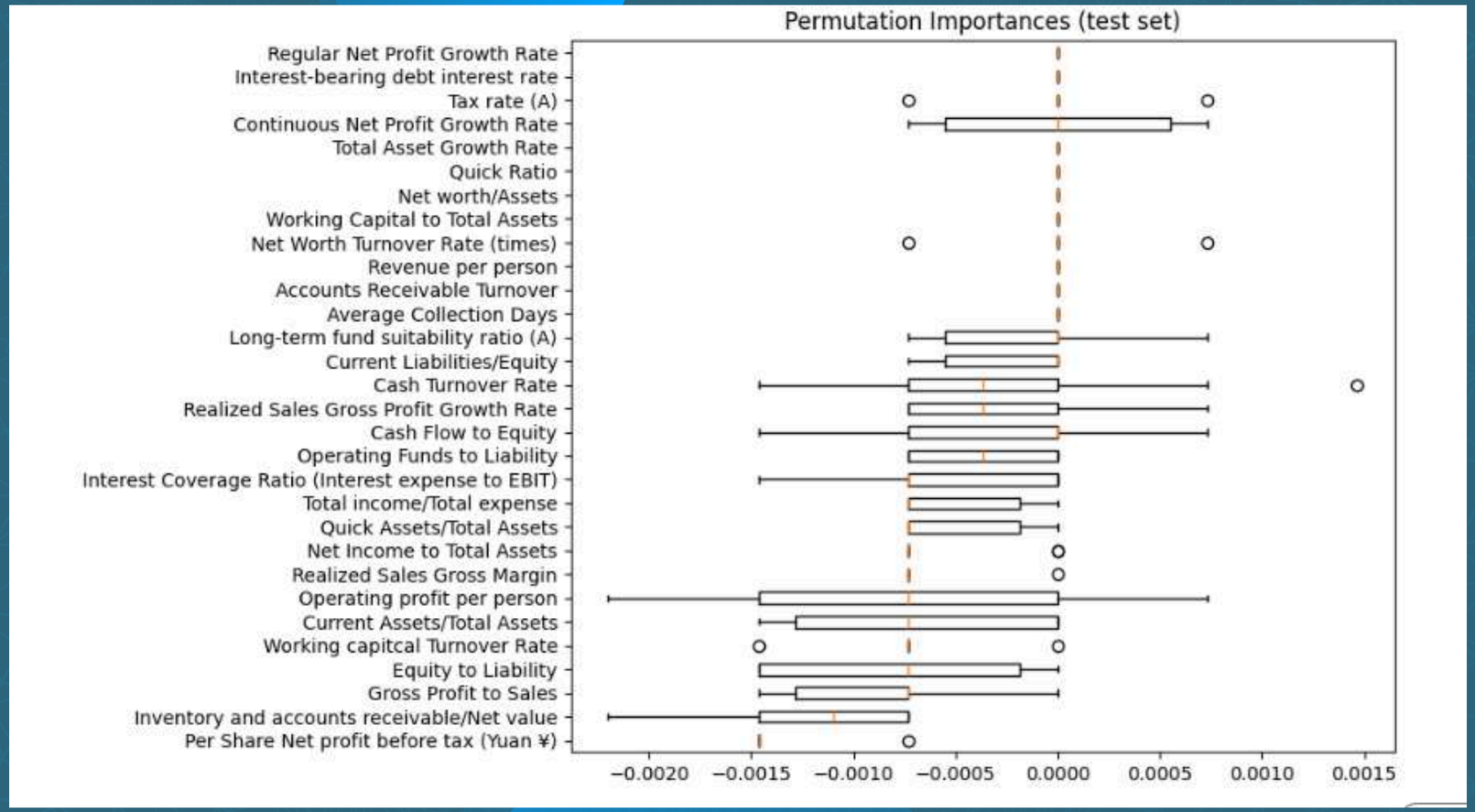


3- LIME



Feature	Value
Operating Profit Per Share (Yuan ¥)	-0.77
Interest Coverage Ratio (Interest expense to EBIT)	-0.02
Borrowing dependency	-0.31
Persistent EPS in the Last Four Seasons	-0.79
Net Income to Stockholder's Equity	-0.14
Operating profit per person	-0.34
Quick Assets/Current Liability	-0.02
Cash Flow to Liability	-0.09
Net Value Growth Rate	-0.01
Cash/Total Assets	-0.55

4- PERMUTATION_IMPORTANCE



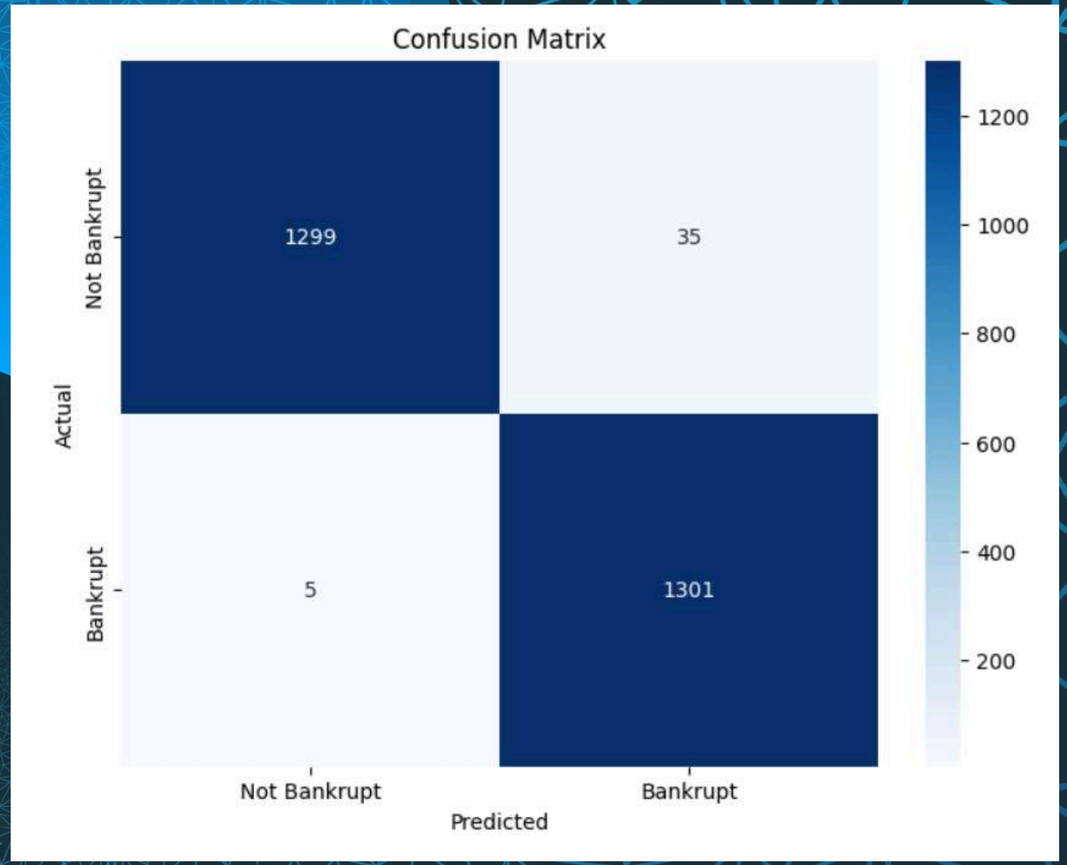
RANDOM FOREST RESULTS

Evaluate the model

```
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.9848484848484849
```

```
Precision: 0.973802395209580 8
Recall: 0.996171516079632 5
F1-score: 0.984859954579863 8
```

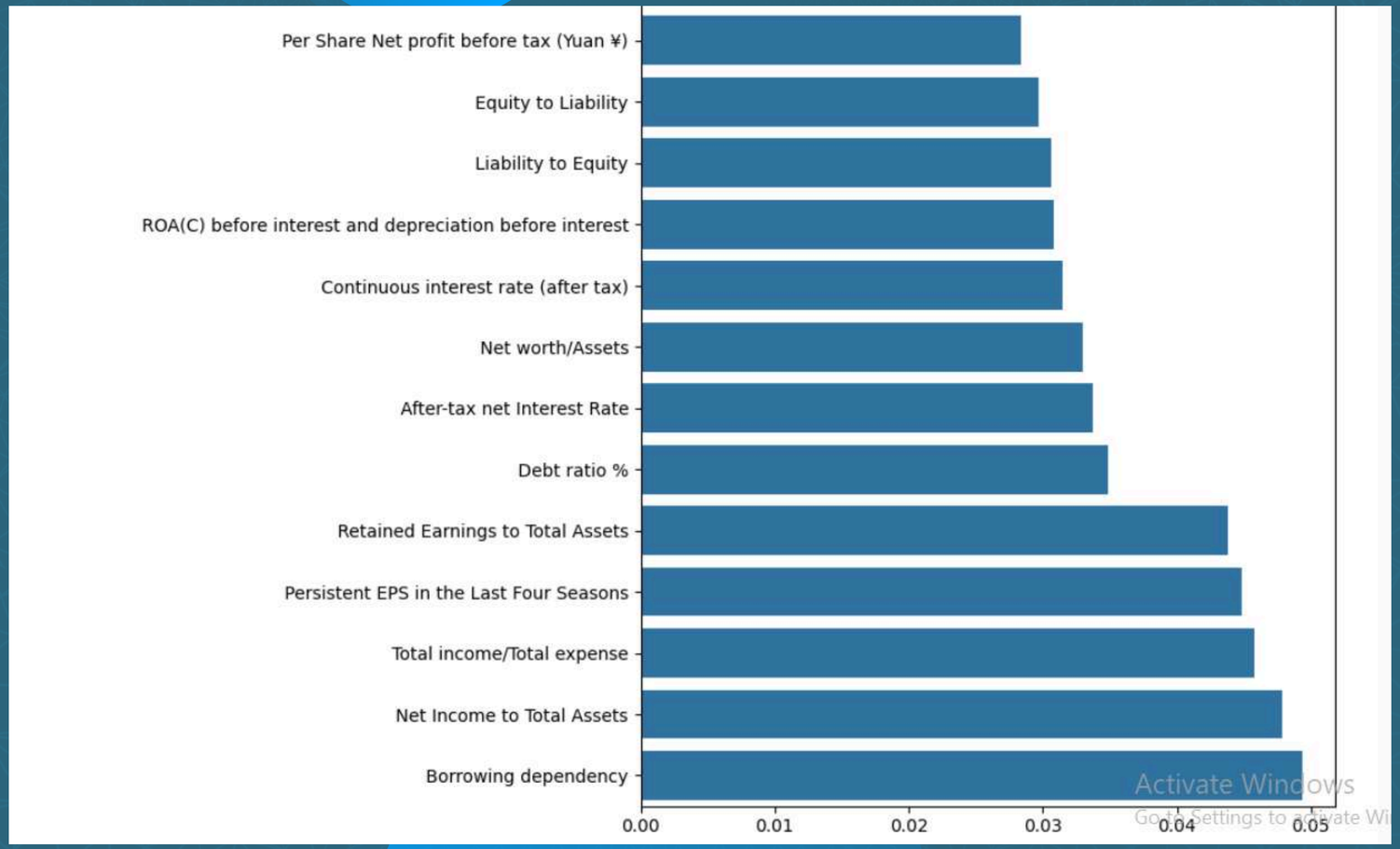


EXPLAINABILITY TECHNIQUES RANDOM FOREST :

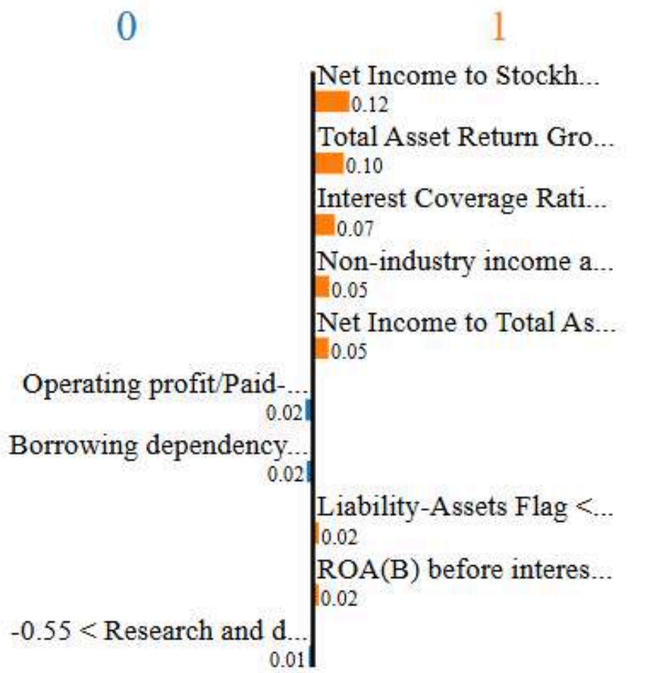
1-Feature Importance

2- LIME

I- FEATURE IMPORTANCE



2- LIME



Feature	Value
Net Income to Stockholder's Equity	-0.14
Total Asset Return Growth Rate Ratio	-0.09
Interest Coverage Ratio (Interest expense to EBIT)	-0.02
Non-industry income and expenditure/revenue	-0.03
Net Income to Total Assets	-1.05
Operating profit/Paid-in capital	-0.77
Borrowing dependency	-0.31
Liability-Assets Flag	-0.03
ROA(B) before interest and depreciation after tax	-0.89
Research and development expense rate	0.42

MACHINE LEARNING MODELS

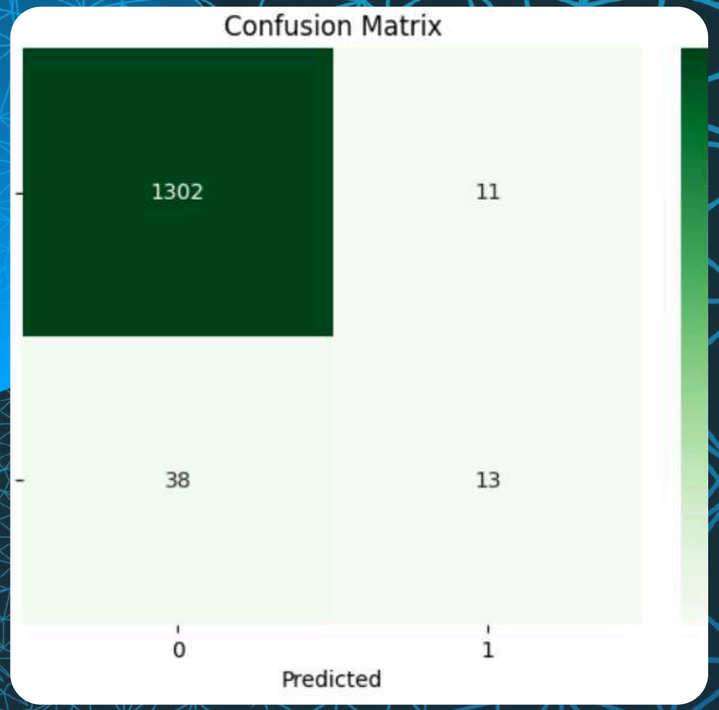
1- XGBOOST (EXTREME GRADIENT BOOSTING) IS A POWERFUL AND SCALABLE IMPLEMENTATION OF GRADIENT BOOSTING THAT BUILDS DECISION TREES SEQUENTIALLY TO MINIMIZE PREDICTION ERRORS, MAKING IT HIGHLY EFFECTIVE FOR STRUCTURED DATA AND COMPETITION-WINNING MODELS.

2- LIGHTGBM (LIGHT GRADIENT BOOSTING MACHINE) IS A HIGH-PERFORMANCE GRADIENT BOOSTING FRAMEWORK. IT USES HISTOGRAM-BASED ALGORITHMS AND GROWS TREES LEAF-WISE, ALLOWING IT TO TRAIN FASTER AND HANDLE LARGE-SCALE DATASETS MORE EFFICIENTLY THAN TRADITIONAL BOOSTING METHODS.

3- SVM (SUPPORT VECTOR MACHINE) IS USED FOR CLASSIFICATION TASKS THAT FIND THE OPTIMAL HYPERPLANE TO SEPARATE DATA INTO CLASSES. IT WORKS WELL FOR BOTH LINEAR AND NON-LINEAR DATA USING KERNEL TRICKS AND IS PARTICULARLY EFFECTIVE IN HIGH-DIMENSIONAL SPACES.

XGBOOST (EXTREME GRADIENT BOOSTING)

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1313
1	0.50	0.24	0.32	51
accuracy			0.96	1364
macro avg	0.74	0.61	0.65	1364
ighted avg	0.95	0.96	0.96	1364



Accuracy: 0.96. It makes the right prediction on most of the data.

Precision: 0.50 Model predicts a “1” class, 0.97 predicts a “0” class.

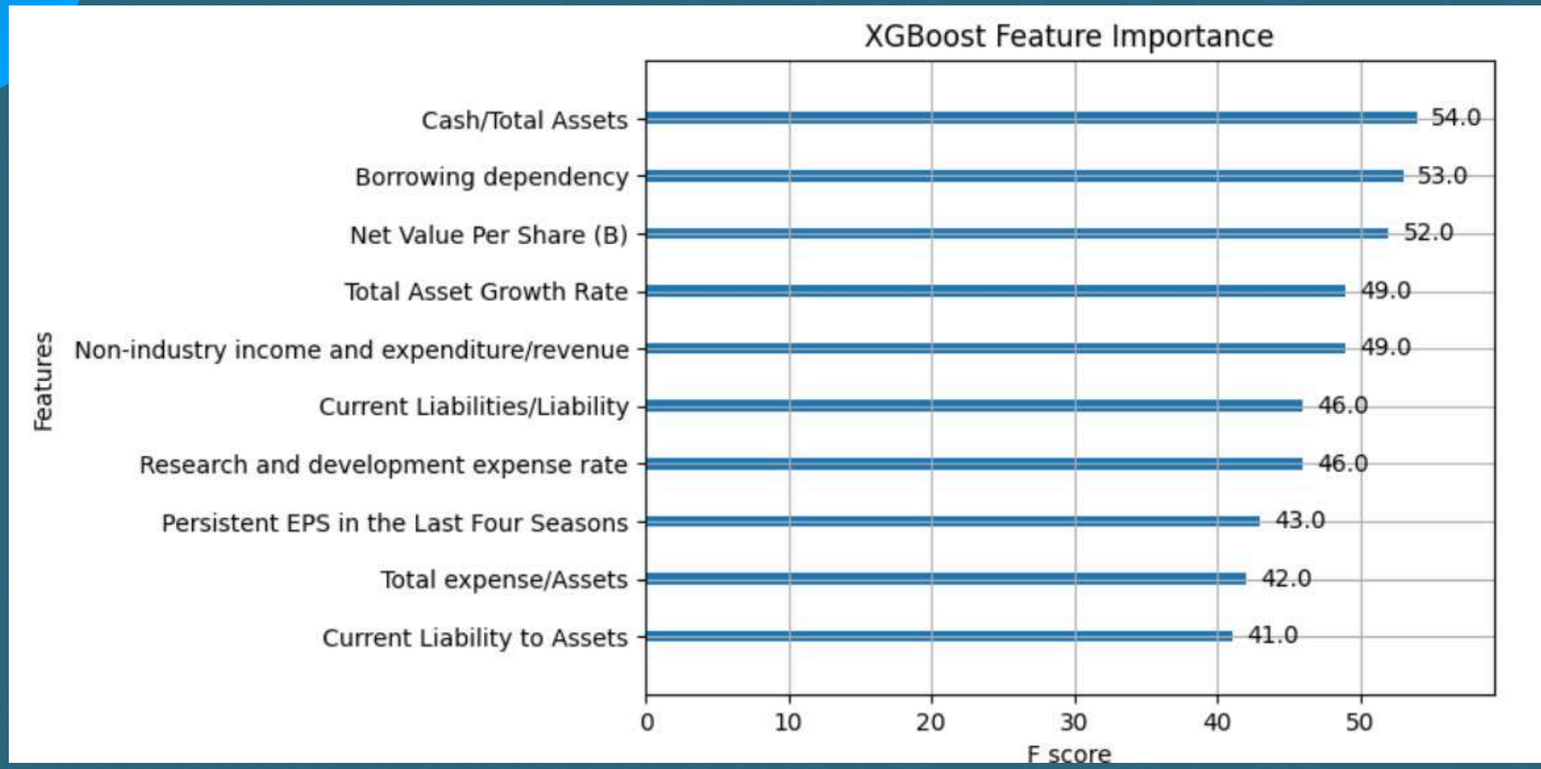
Recall: 0.24 Model predicts a “1” class, 0.99 predicts a “0” class.

EXPLAINABILITY TECHNIQUES XGBOOST

1- Feature Importance
2- LIME
3- PDP
4- Shap
5- ICE

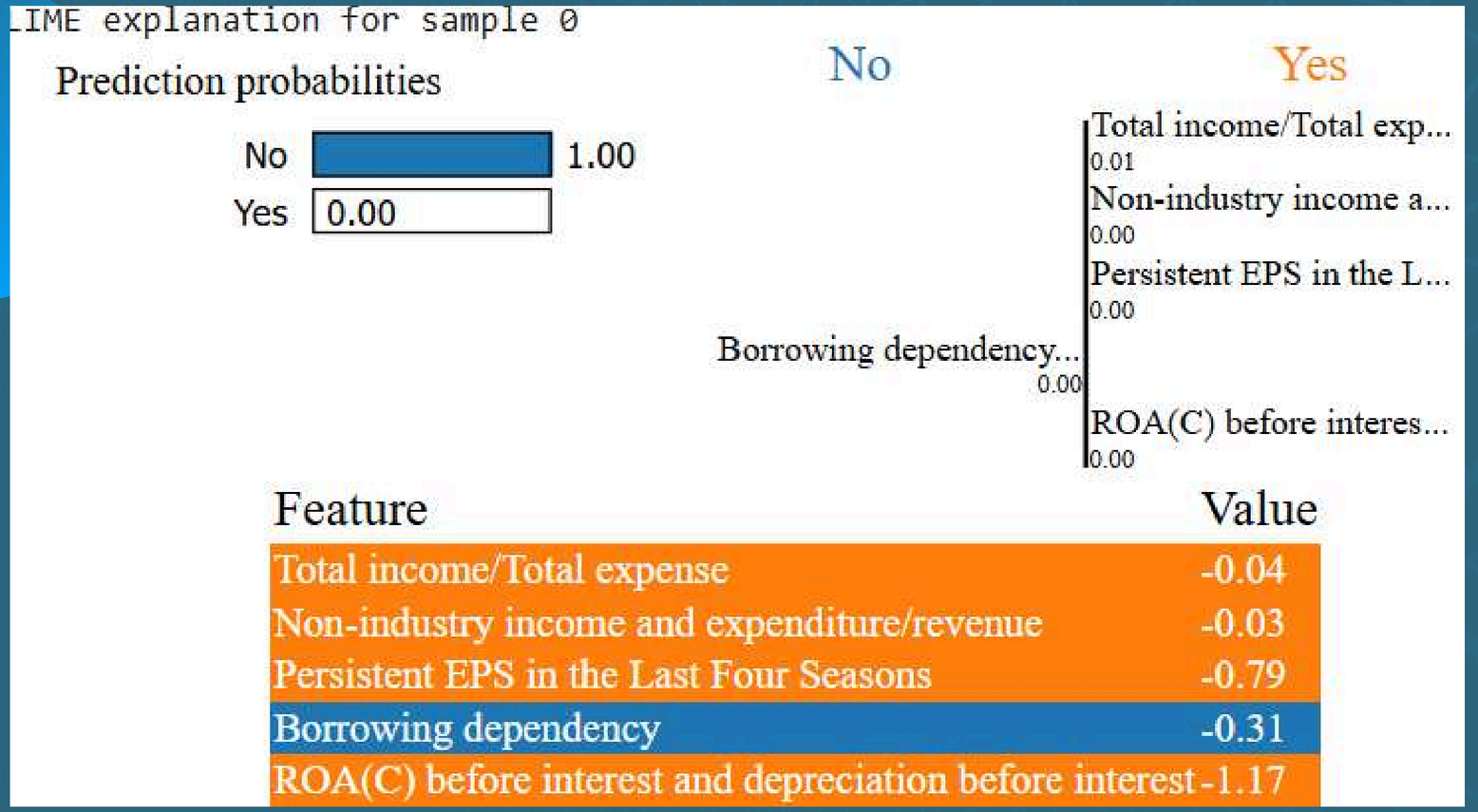
I- FEATURE IMPORTANCE

The feature importance plot displays the most greater top 10 features from the data that affect the performance of the model and have a powerful impact on the training and validation



2- LIME

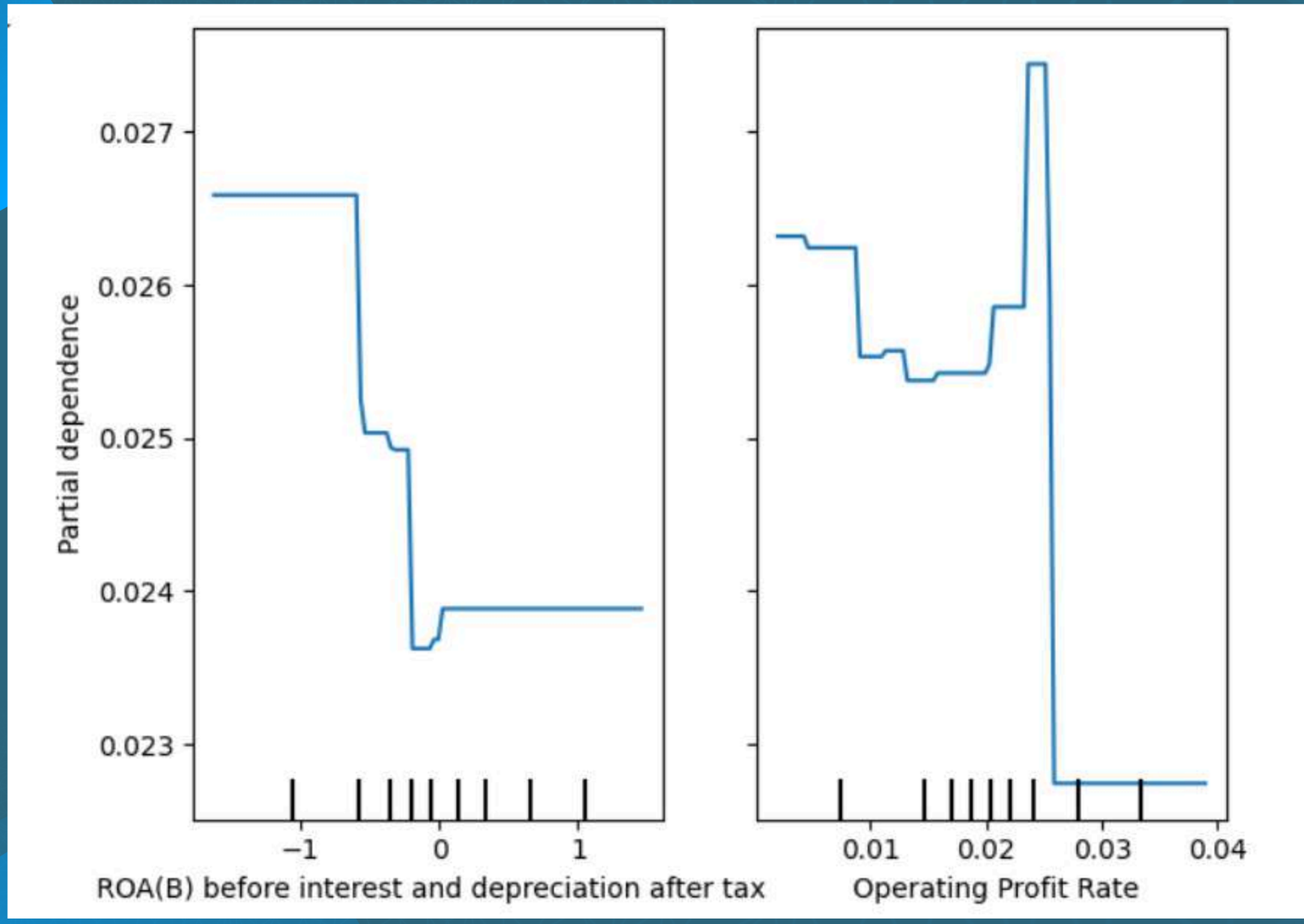
LIME is used to interpret predictions made by any black-box machine learning model, it approximates the complex model locally around the prediction.



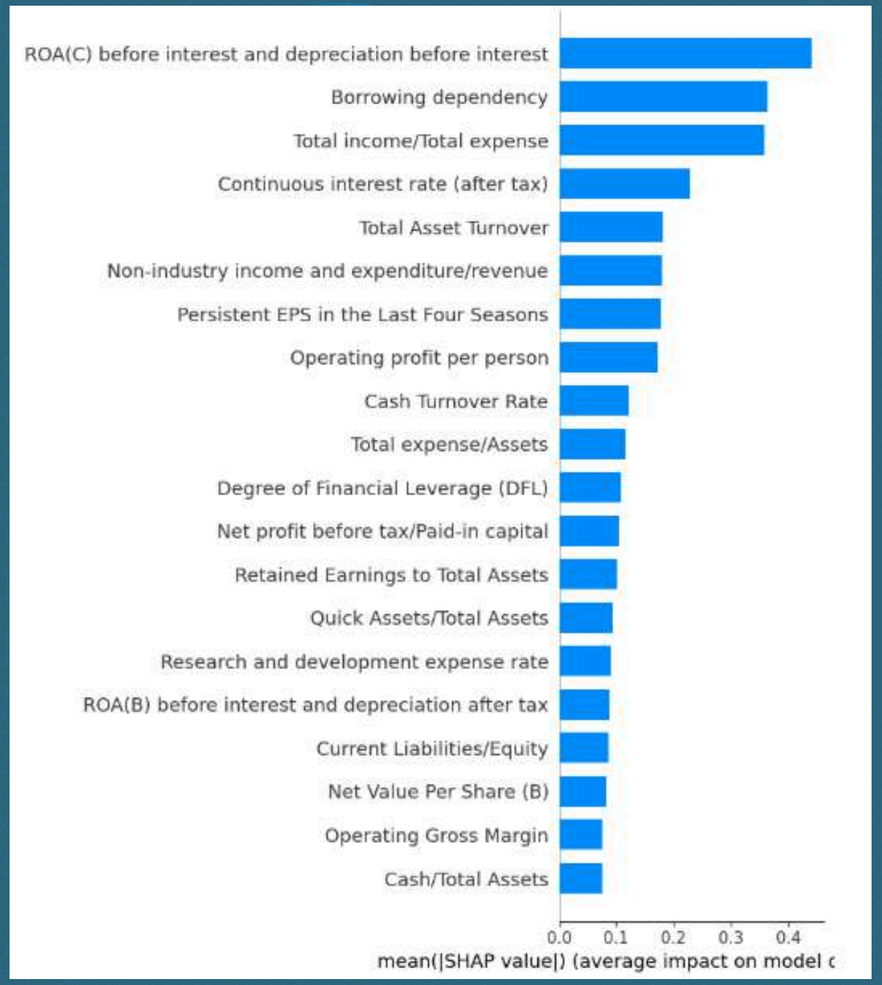
For sample feature 0, even though there are negative signs like high total income and borrowing dependency, it has good profits as the model predicts with 99% certainty that the company is not bankrupt.

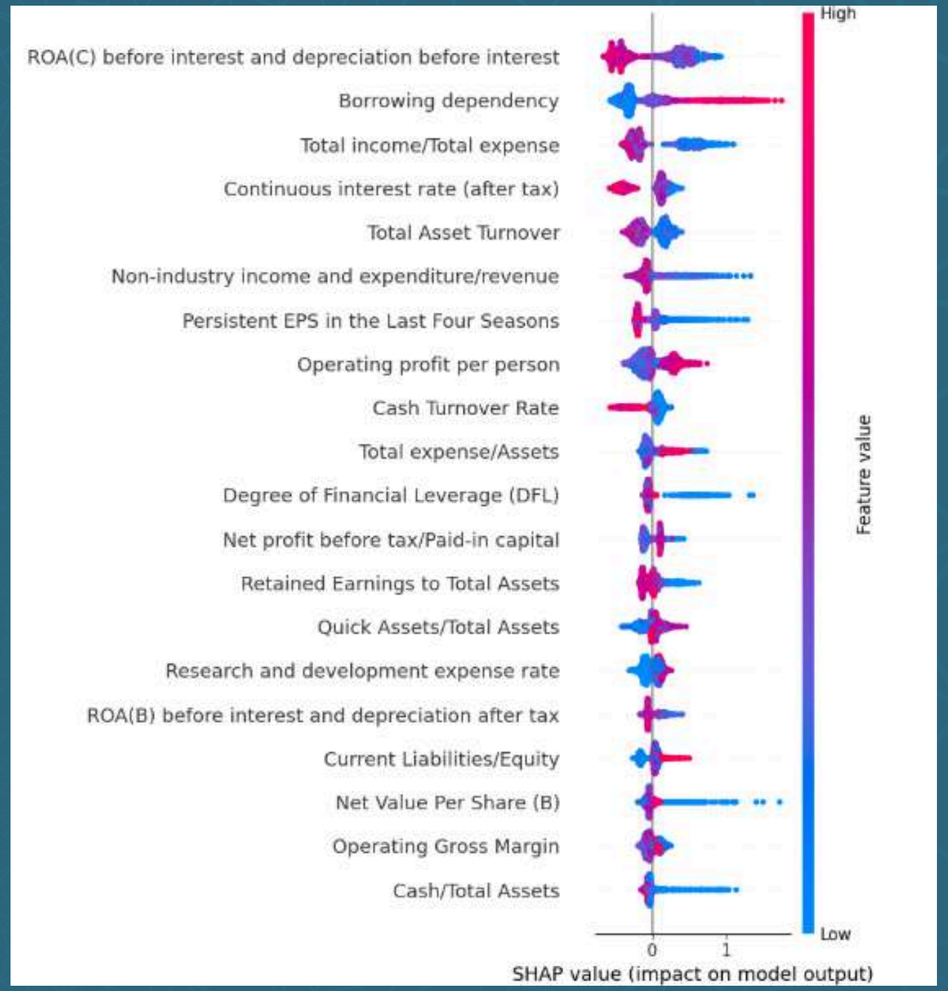
3- PDP

PDP is a model-agnostic interpretability technique that shows the marginal effect of one or more features on the predicted outcome.
It visualizes for these two features value how influence based on the model's prediction, while averaging all other features.



4- SHAP

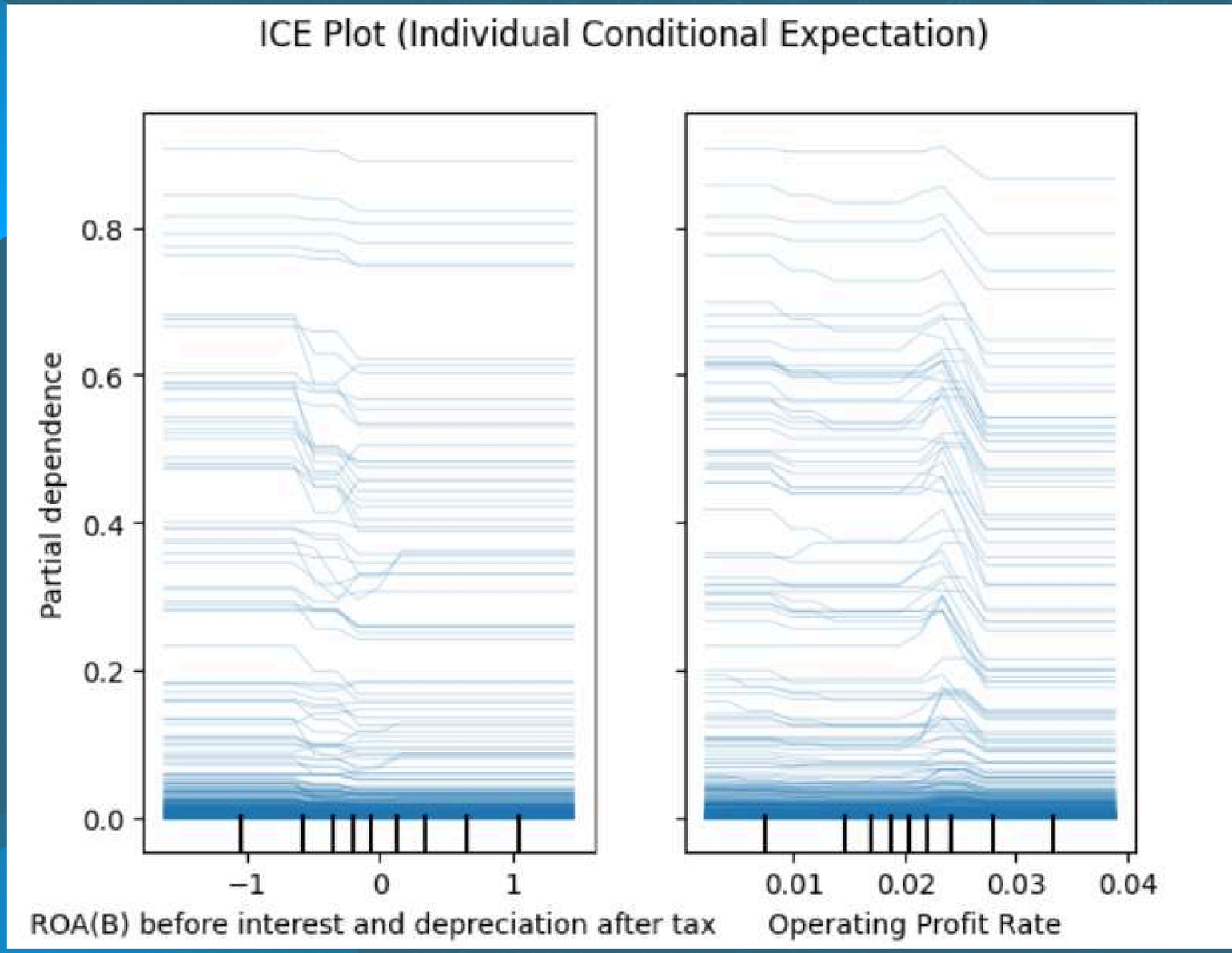




The Summary plot of shap shows the mean shap value as it's the absolute average impact on the model, as here the top Roa(c) feature. The beeswarm plot displays the same idea but with non-absolute values and range high-low

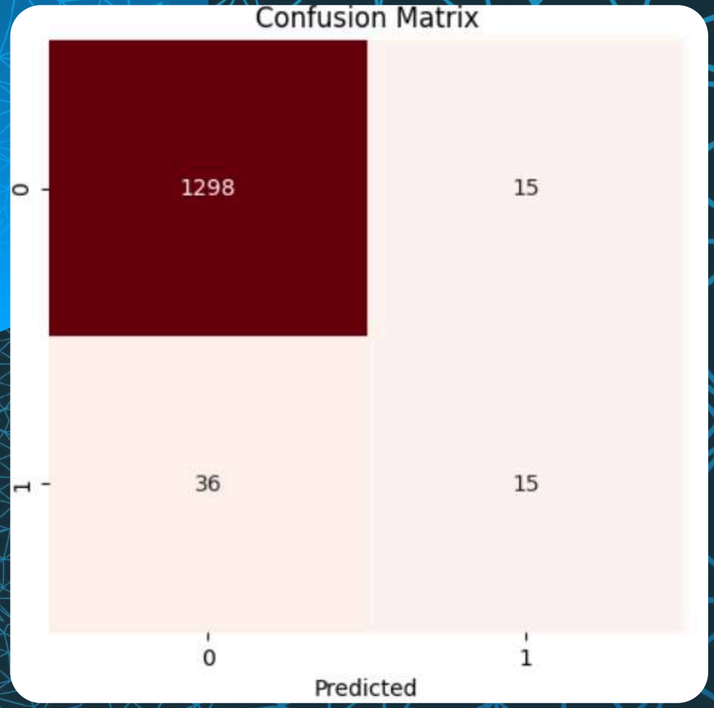
5- ICE

ICE plots show how a model’s prediction changes for an individual data instance as a single feature varies, keeping all other features fixed. provides insight into how different individuals are affected by a feature for each value, a line in the plot.



LIGHTGBM
(LIGHT GRADIENT BOOSTING MACHINE)

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1313
1	0.50	0.29	0.37	51
accuracy			0.96	1364
macro avg	0.74	0.64	0.68	1364
weighted avg	0.96	0.96	0.96	1364



Accuracy: 0.96. It makes the right prediction on most of the data.

Precision: 0.50 Model predicts a “1” class, 0.97 predicts a “0” class.

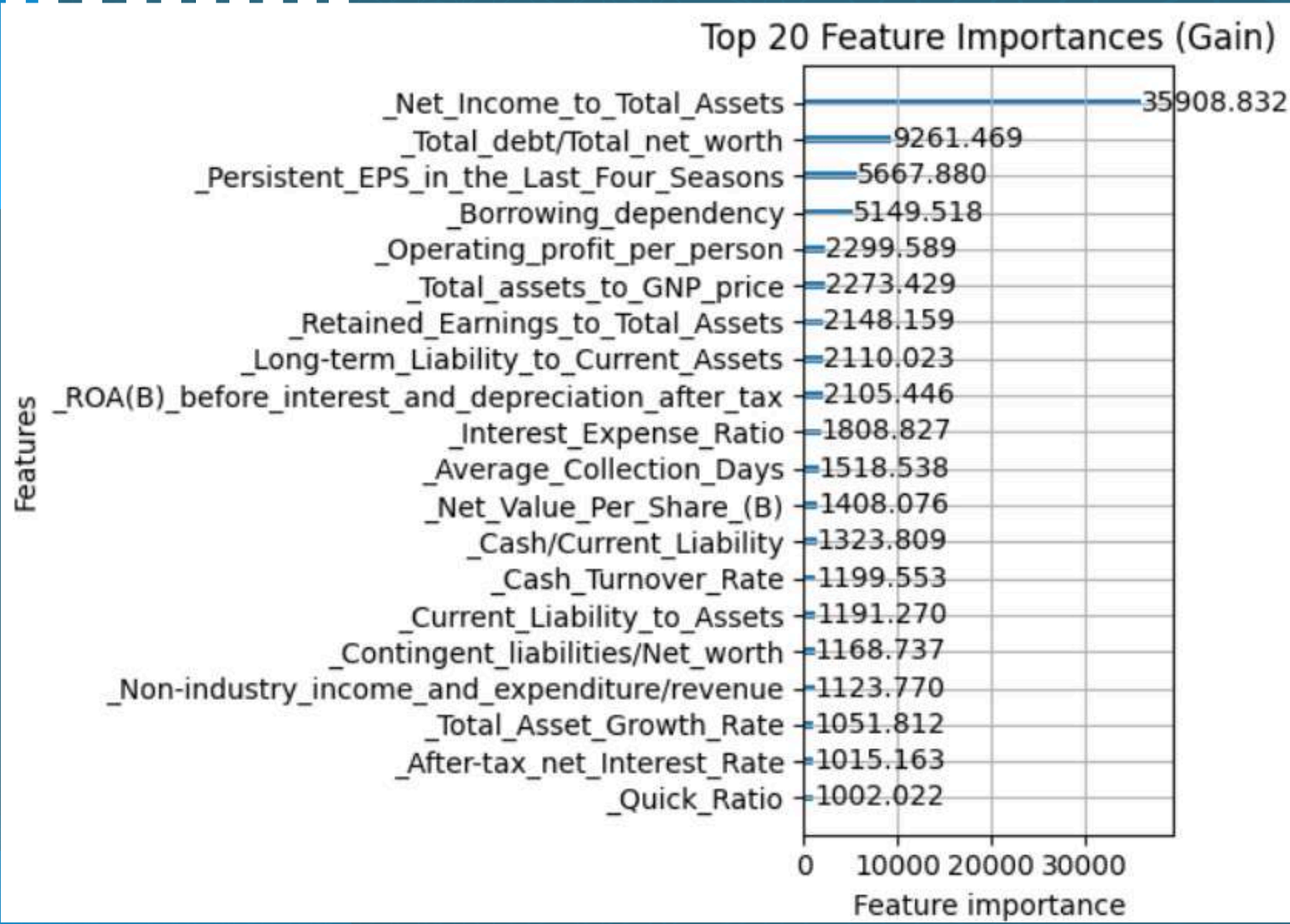
Recall: 0.29 Model predicts a “1” class, 0.99 predicts a “0” class.

EXPLAINABILITY TECHNIQUES LIGHTGBM

1- Feature Importance
2- LIME
3- PDP
4- Shap
5- ICE

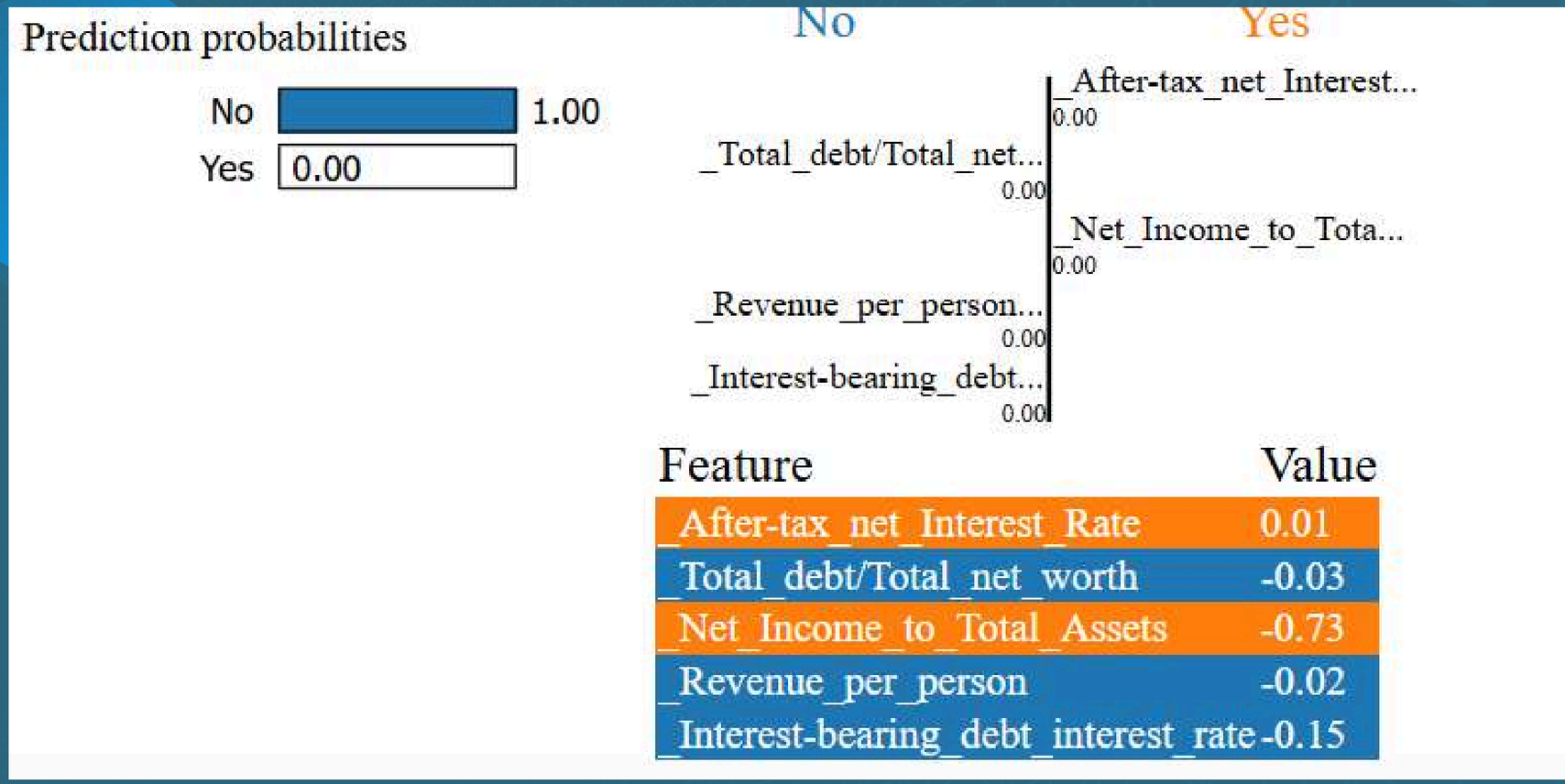
I- FEATURE IMPORTANCE

The feature importance plot displays the most greater top 10 features from the data that affect the performance of the model and have a powerful impact on the training and validation



2- LIME

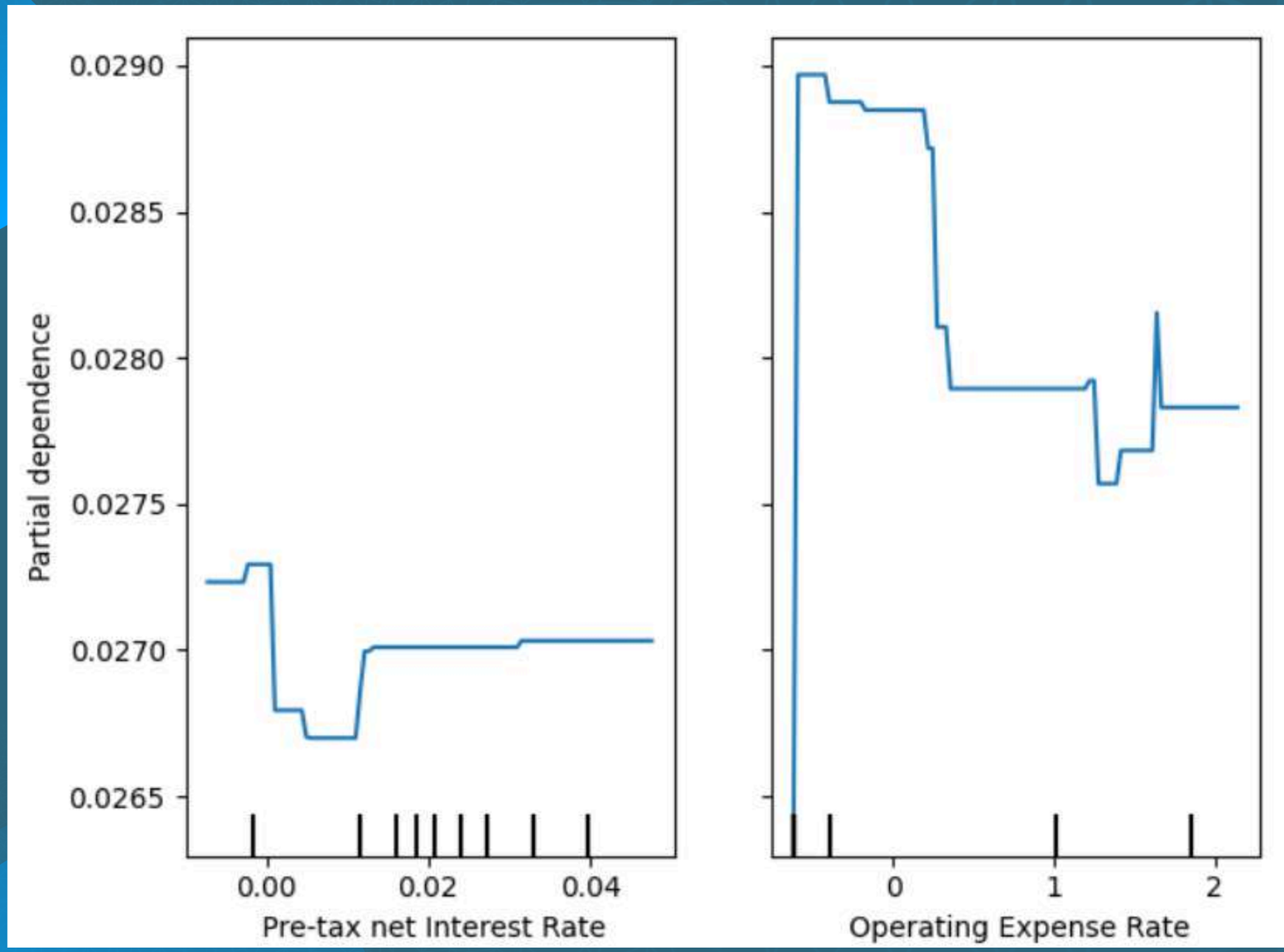
LIME is used to interpret predictions made by any black-box machine learning model, it approximates the complex model locally around the prediction.



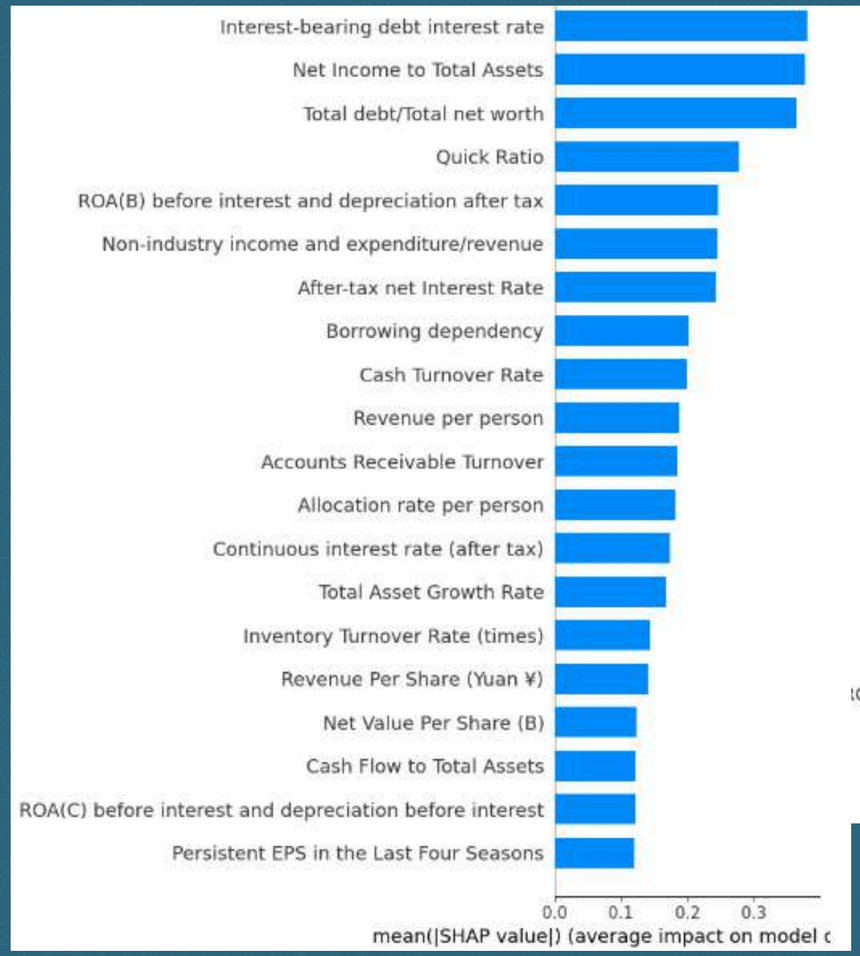
For sample feature, even though there are negative signs net income and after tax net, it has good profits as the model predicts with 99% certainty that the company is not bankrupt.

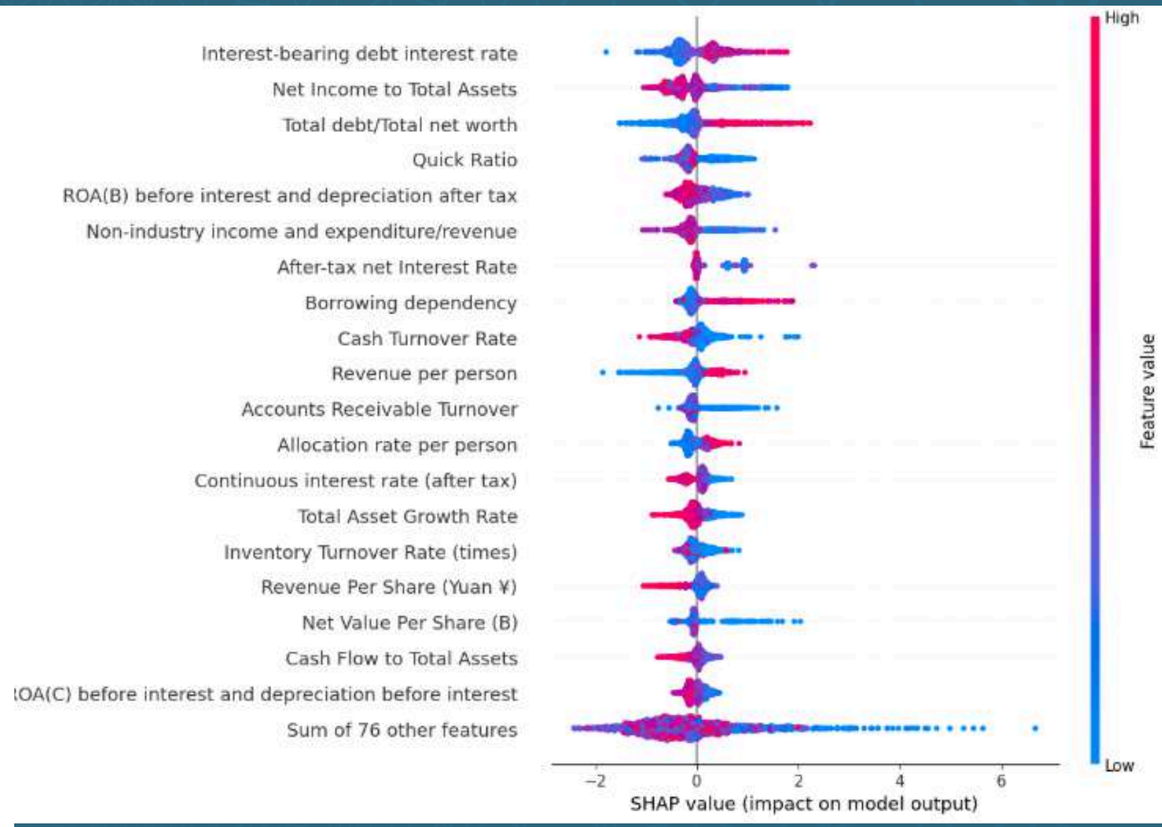
3- PDP

PDP is a model-agnostic interpretability technique that shows the marginal effect of one or more features on the predicted outcome.
It visualizes for these two features value how influence based on the model's prediction, while averaging all other features.



4- SHAP

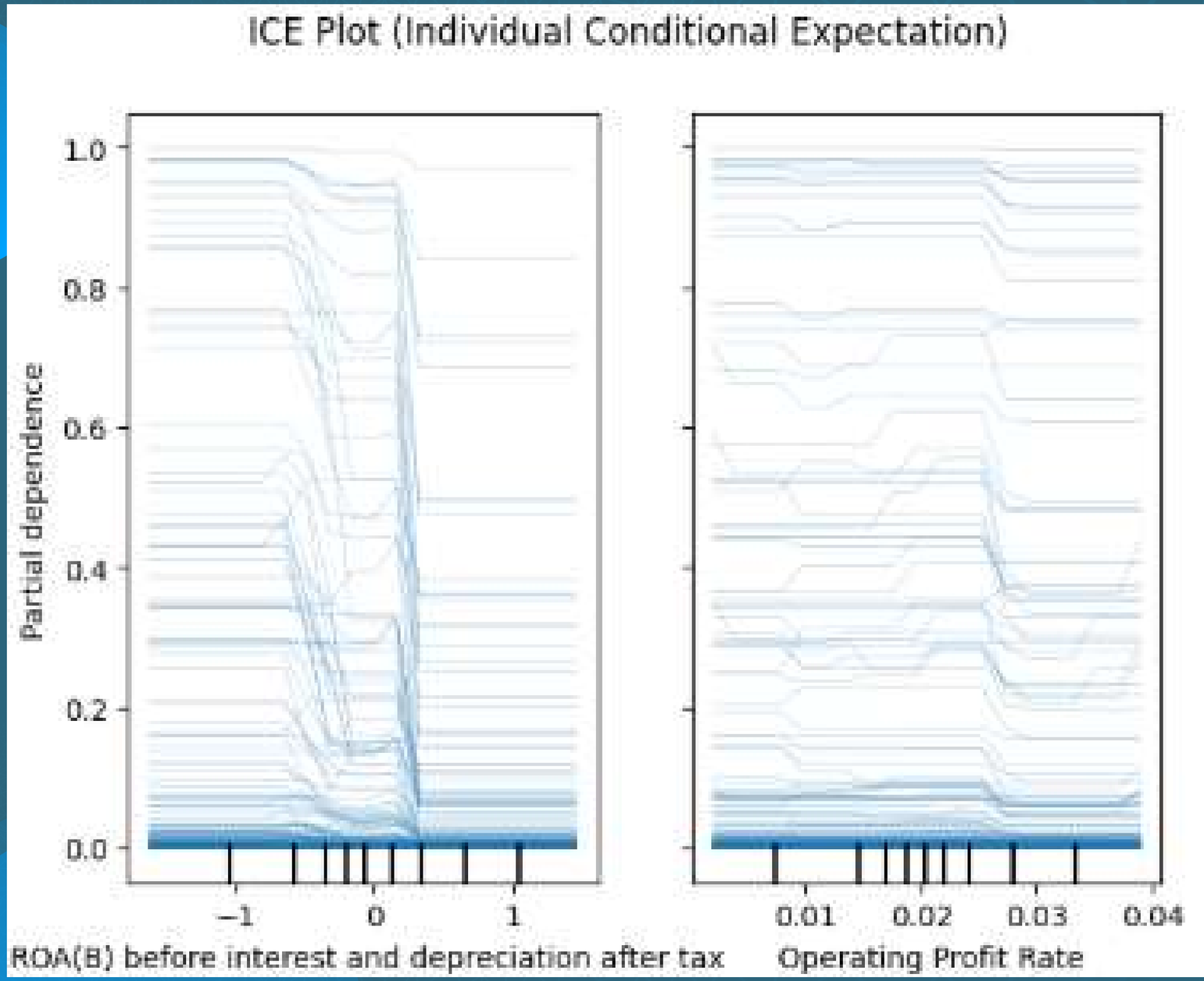




The Summary plot of shap shows the mean shap value as it's the absolute average impact on the model, as here the top interest bearing dept feature. The beeswarm plot displays the same idea but with non-absolute values and range high-low

5- ICE

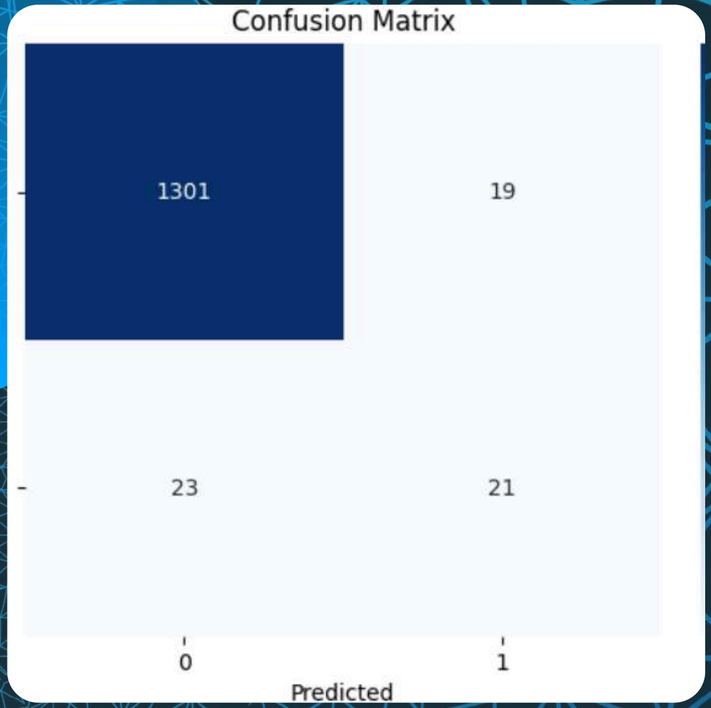
ICE plots show how a model’s prediction changes for an individual data instance as a single feature varies, keeping all other features fixed. provides insight into how different individuals are affected by a feature for each value, a line in the plot.



SVM
(SUPPORT VECTOR MACHINE)

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	1320
1	0.53	0.48	0.50	44
accuracy			0.97	1364
macro avg	0.75	0.73	0.74	1364
weighted avg	0.97	0.97	0.97	1364



Accuracy: 0.97. It makes the right prediction on most of the data.
Precision: 0.53 Model predicts a “1” class, 0.8 predicts a “0” class.
Recall: 0.48 Model predicts a “1” class, 0.99 predicts a “0” class.

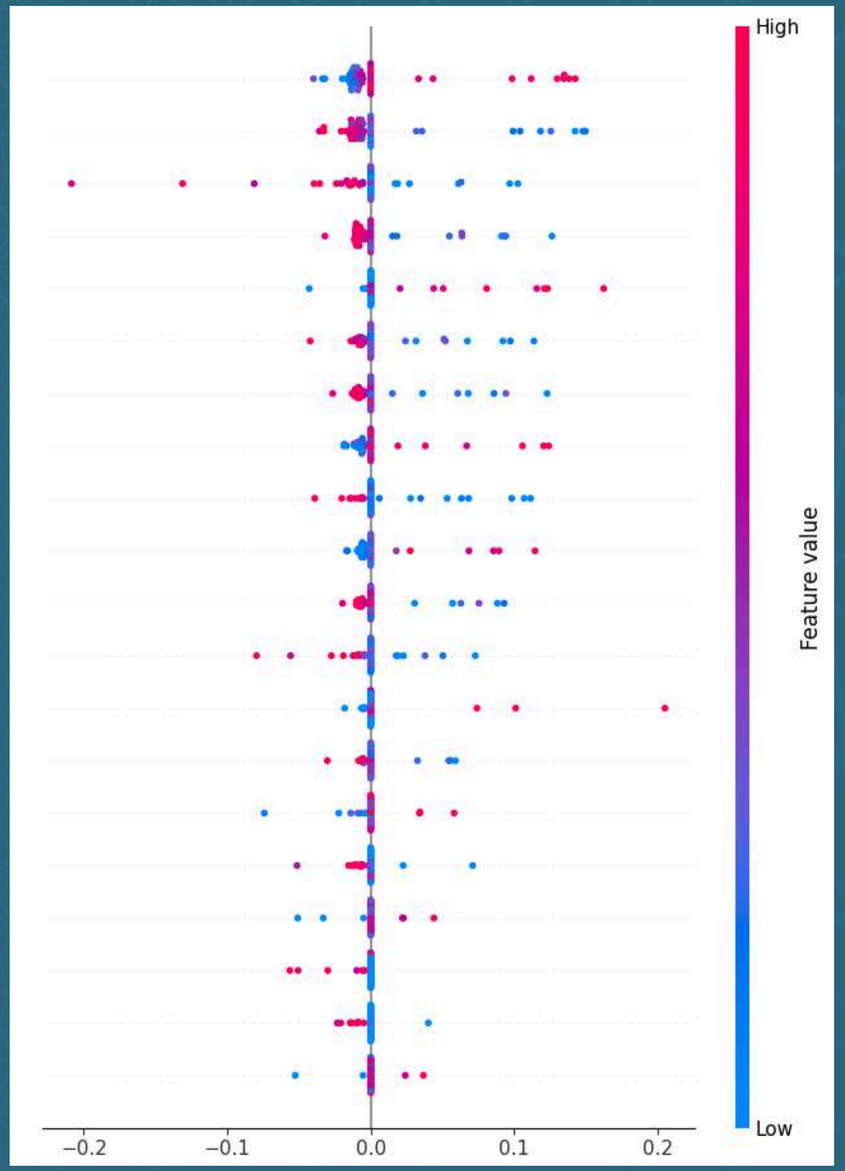
1- SVM PERMUTATION IMPORTANCE

The feature importance plot displays the most greater top 10 features from the data that affect the performance of the model and have a powerful impact on the training and validation

Permutation Importance (mean decrease in F1):	
Net Income to Total Assets	0.347082
ROA(C) before interest and depreciation before interest	0.324569
Net Value Per Share (B)	0.320366
Persistent EPS in the Last Four Seasons	0.274655
ROA(A) before interest and % after tax	0.268201
After-tax Net Profit Growth Rate	0.267647
ROA(B) before interest and depreciation after tax	0.246533
Regular Net Profit Growth Rate	0.174684
Total debt/Total net worth	0.141290
Net Value Per Share (C)	0.140326

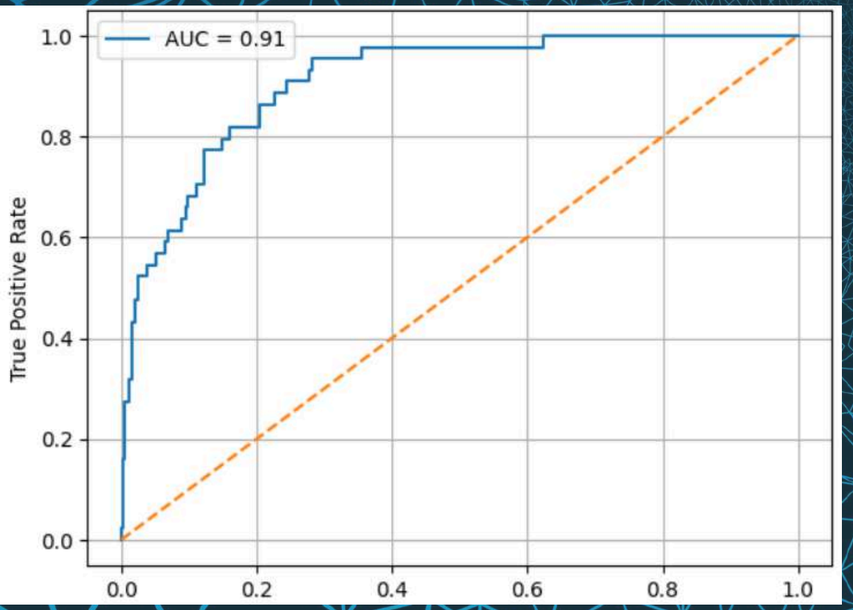
4- SVM SHAP

The Summary plot of shap shows the mean shap value as it's the absolute average impact on the model, as here the top interest bearing dept feature. The beeswarm plot displays the same idea but with non-absolute values and range high-low

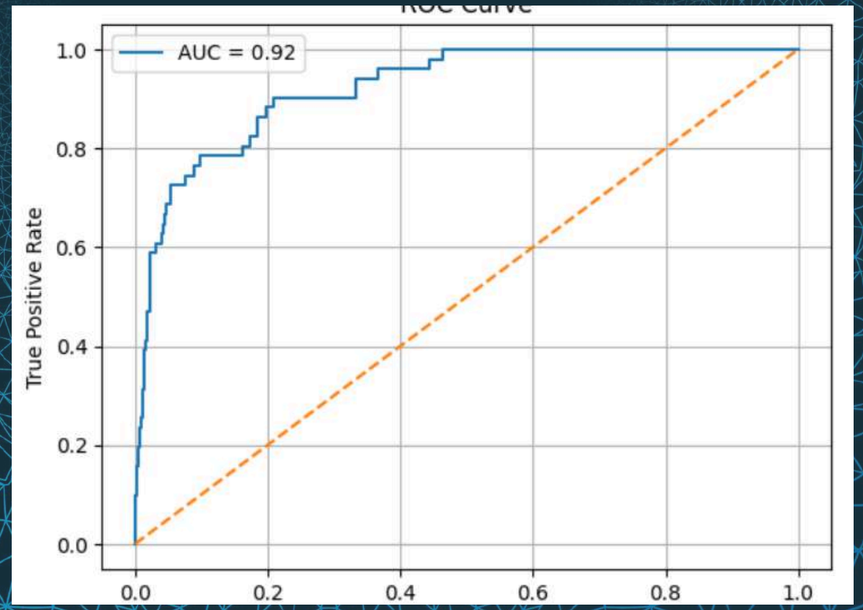


ROC Comparison

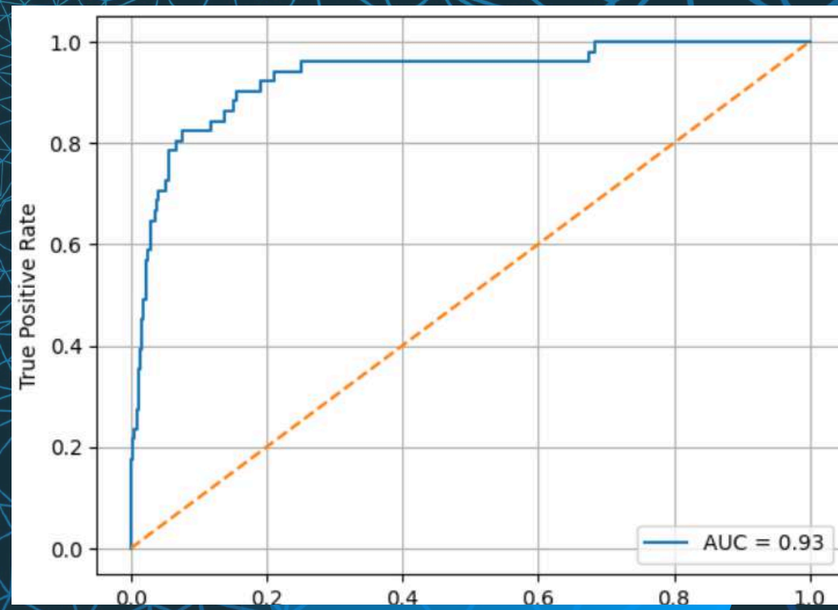
SVM ROC PLOT



XGBOOST ROC PLOT



LIGHTGBM ROC PLOT



MACHINE LEARNING MODELS

1

We applied an XGBoost classifier to the oversampled training data for robust and efficient classification. After training, we evaluated its performance using a classification report and confusion matrix, which offered a detailed view of accuracy and class-level prediction quality.

2

We trained a Logistic Regression model on the oversampled training set to address class imbalance, using selected features and a fixed random state for reproducibility. Its performance was evaluated using a classification report and a confusion matrix, providing clear insights into prediction accuracy and class-wise performance

3

We trained a neural network using an MLPClassifier with two hidden layers to capture complex patterns in the oversampled data. Its performance was evaluated using a classification report and confusion matrix, providing insight into prediction accuracy and class-wise results.

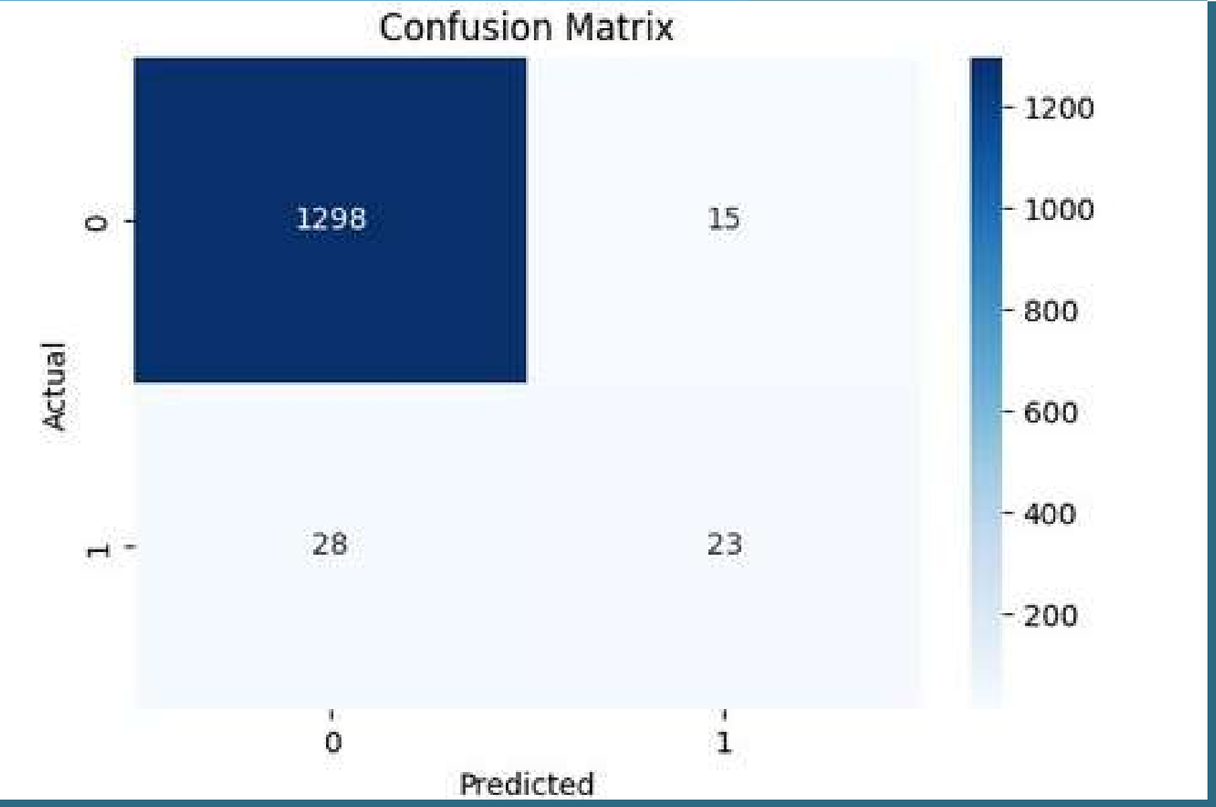
MODEL 01

Accuracy Score: 0.9684750733137831

F1 Score: 0.5168539325842696

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	1313
1	0.61	0.45	0.52	51
accuracy			0.97	1364
macro avg	0.79	0.72	0.75	1364
weighted avg	0.96	0.97	0.97	1364



High overall accuracy (96.8%), driven mainly by strong detection of non-bankrupt companies.

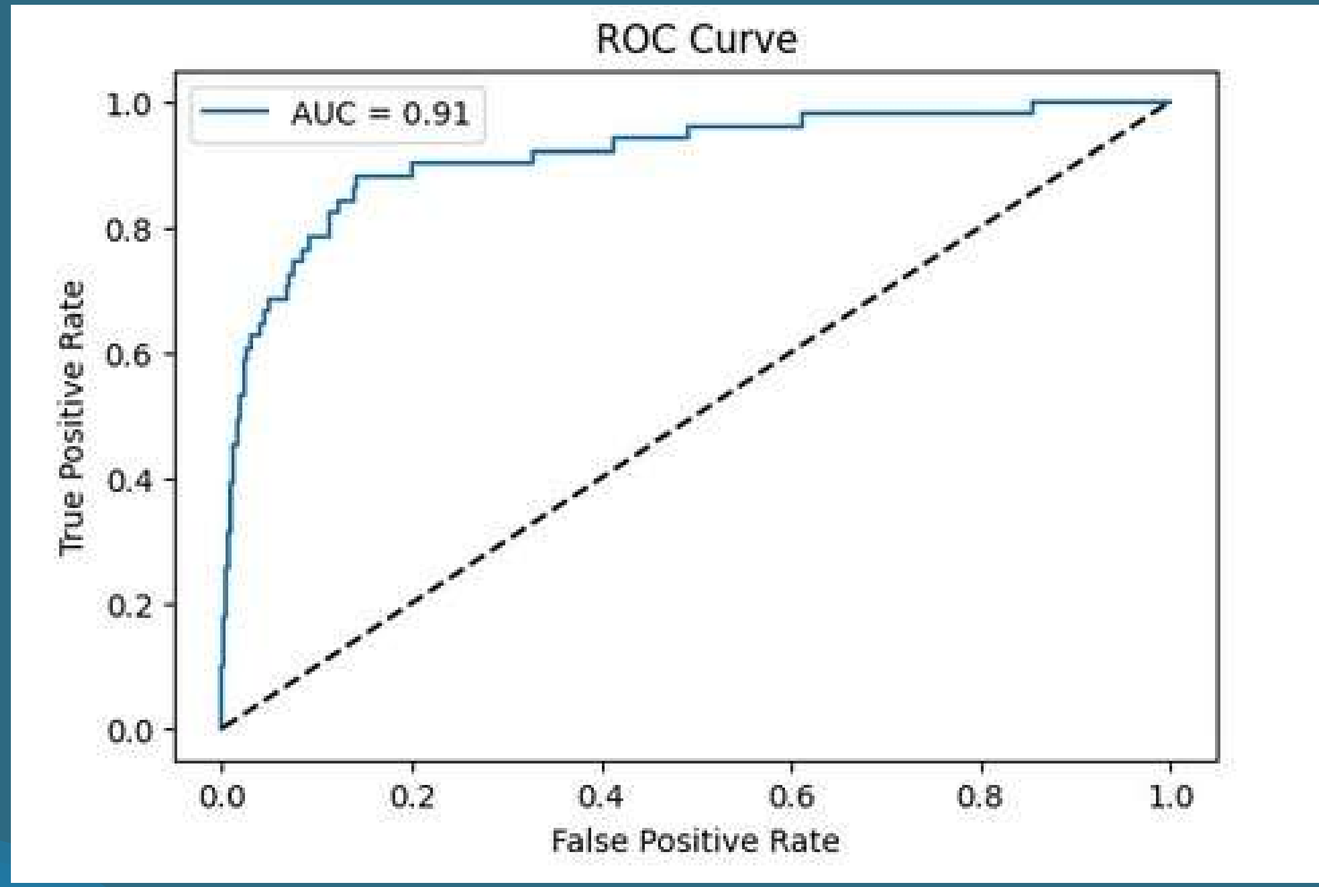
Low recall (45%) for bankrupt class → the model misses many actual bankrupt cases.

Precision for bankrupt class is decent (61%), indicating predictions are somewhat reliable when it does detect bankruptcy.

Model may require further tuning or feature engineering to improve minority class performance.

ROC CURVE

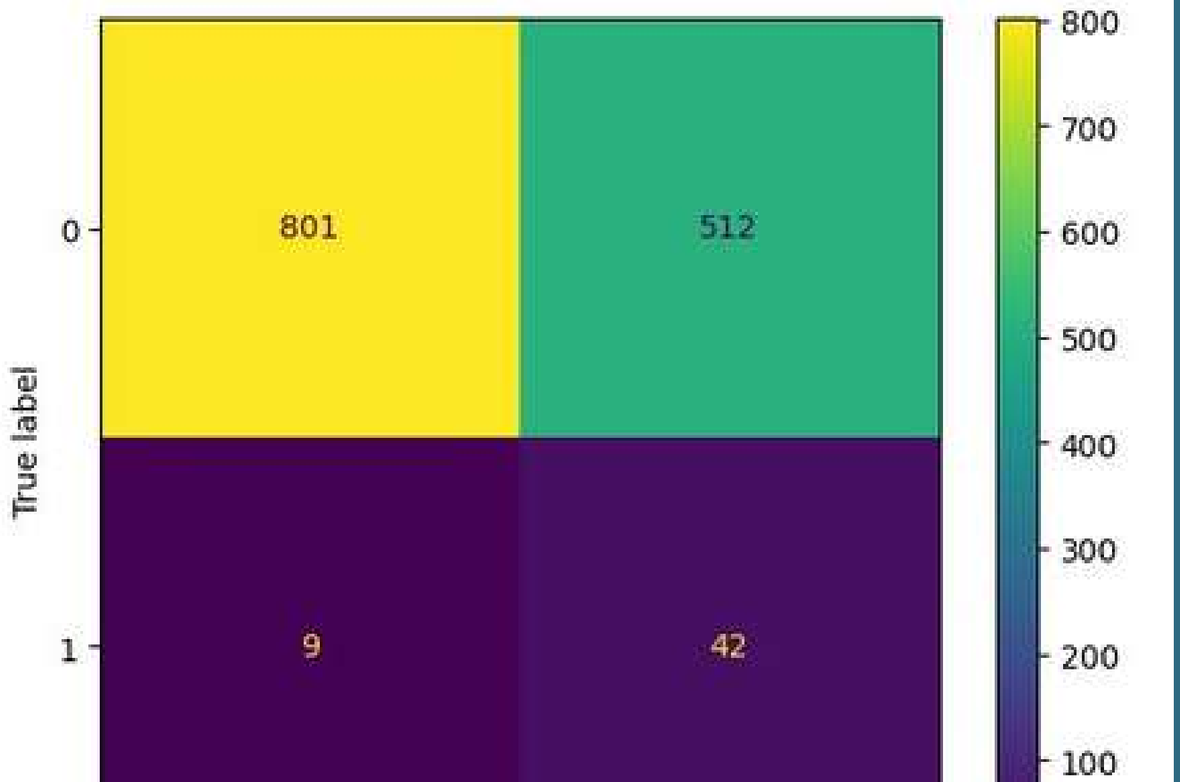
- The image is a histogram showing the distribution of the Net Income to Total Assets Ratio. Most values cluster between 0.75 and 0.9, indicating that the majority of entities have high profitability relative to their assets. This is a data distribution plot, not an ROC curve, which is used for evaluating classification models.



MODEL 02

Report:

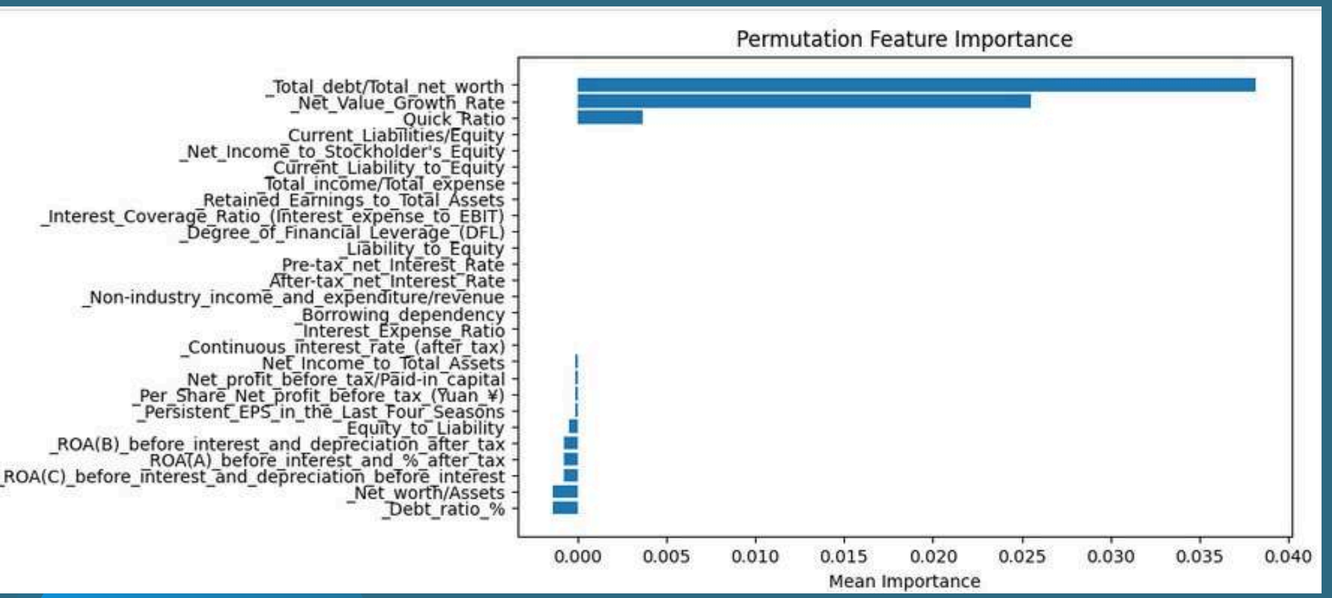
	precision	recall	f1-score	support
	0.99	0.61	0.75	1313
	0.08	0.82	0.14	51
			0.62	1364
	0.53	0.72	0.45	1364
	0.95	0.62	0.73	1364



The model achieved 62% accuracy, performing well on the majority class (0) with high precision but struggled with the minority class (1), showing low precision despite high recall. This indicates class imbalance affected overall performance.

EXPLAINABILITY TECHNIQUES LOGISTIC REGRESSION MODEL

Permutation Feature Importance



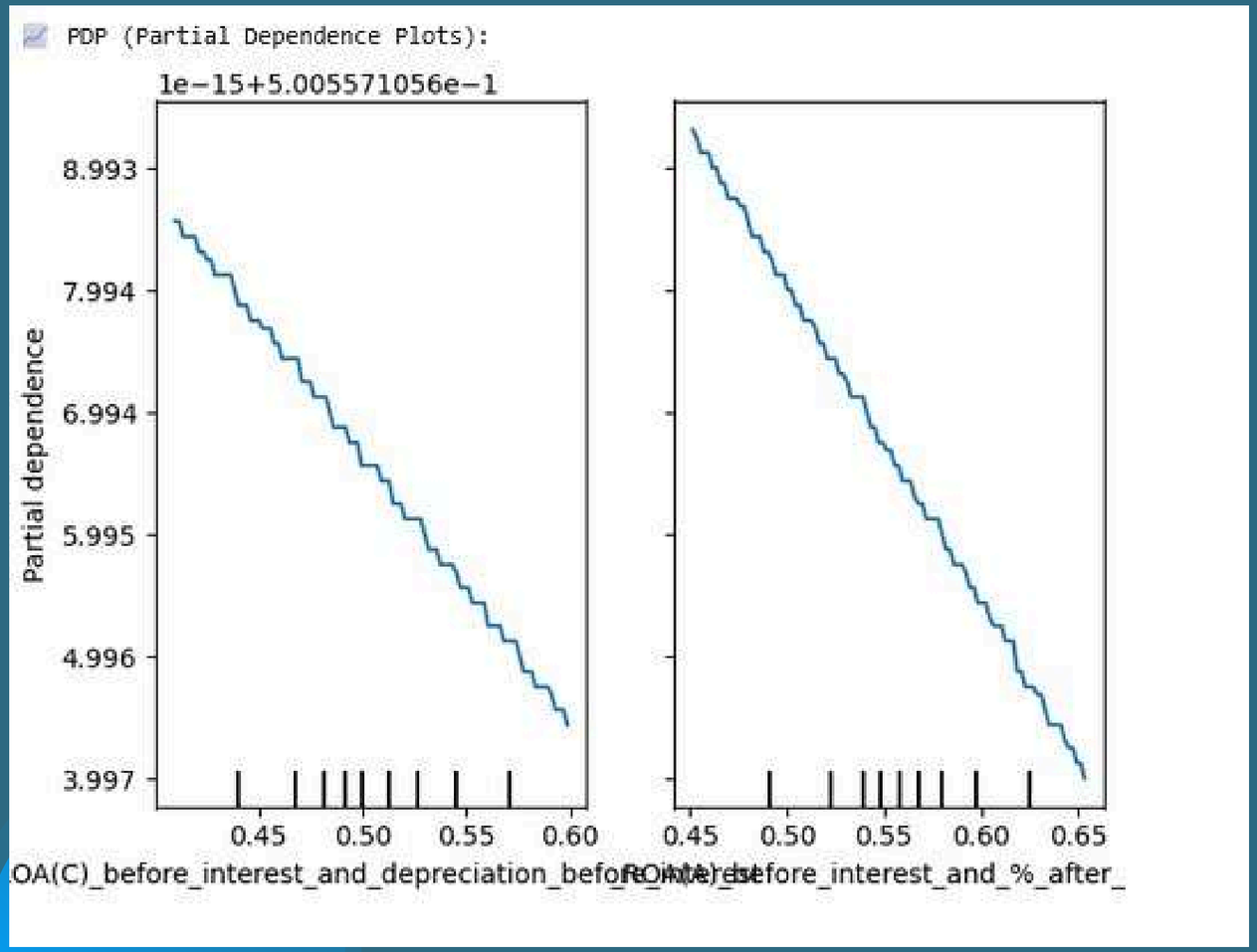
USING PERMUTATION IMPORTANCE ON THE TEST SET, I FOUND THAT THE MOST CRITICAL FEATURES FOR PREDICTING BANKRUPTCY RISK WERE THE TOTAL DEBT TO NET WORTH RATIO, NET VALUE GROWTH RATE, AND QUICK RATIO, WITH THE DEBT-TO-NET-WORTH RATIO HAVING THE HIGHEST IMPACT ON MODEL PERFORMANCE. THESE FEATURES REFLECT A COMPANY'S FINANCIAL STABILITY AND GROWTH, STRONGLY INFLUENCING PREDICTIONS. SECONDARY FEATURES RELATED TO LEVERAGE AND PROFITABILITY ALSO CONTRIBUTED BUT TO A LESSER EXTENT. SEVERAL FEATURES HAD NEAR-ZERO IMPORTANCE, SUGGESTING THEY CAN BE REMOVED TO SIMPLIFY THE MODEL WITHOUT SACRIFICING ACCURACY. OVERALL, THE MODEL CAN BE STREAMLINED BY FOCUSING ON THE TOP 5–7 IMPACTFUL FEATURES.

LIME

Feature	Value
_Quick_Ratio	0.01
_Net_Value_Growth_Rate	0.00
_Debt_ratio_%	0.04
_Equity_to_Liability	0.09
_Total_debt/Total_net_worth	0.00
_Net_worth/Assets	0.96
_Liability_to_Equity	0.28
_ROA(A)_before_interest_and_%_after_tax	0.48
_Interest_Expense_Ratio	0.63
_ROA(C)_before_interest_and_depreciation_before_interest	0.43

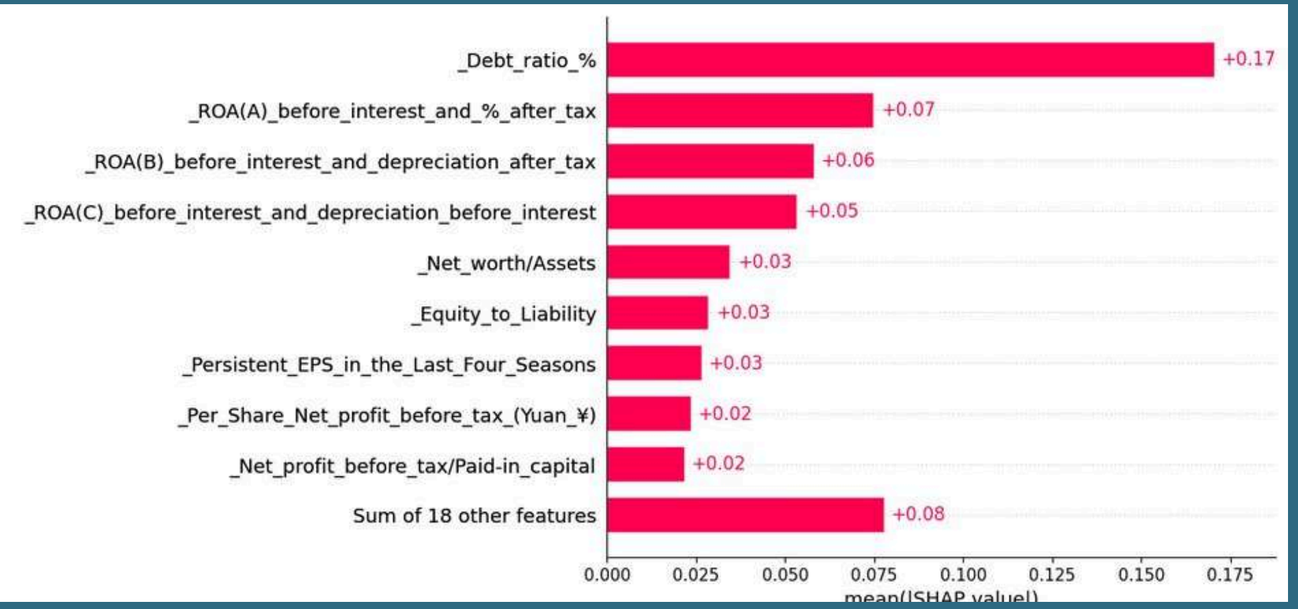
THE LIME EXPLANATION SHOWED THE MODEL WAS 88% CONFIDENT THE COMPANY IS NOT BANKRUPT. KEY FACTORS SUPPORTING THIS PREDICTION INCLUDED A HIGH *ROA (0.50), A **QUICK RATIO OF 0.01* INDICATING SUFFICIENT LIQUIDITY, AND AN *EQUITY-TO-LIABILITY RATIO OF 0.09, ALL SUGGESTING FINANCIAL STABILITY. THE **NET VALUE GROWTH RATE* HAD MINIMAL INFLUENCE ON THE OUTCOME.

PDP



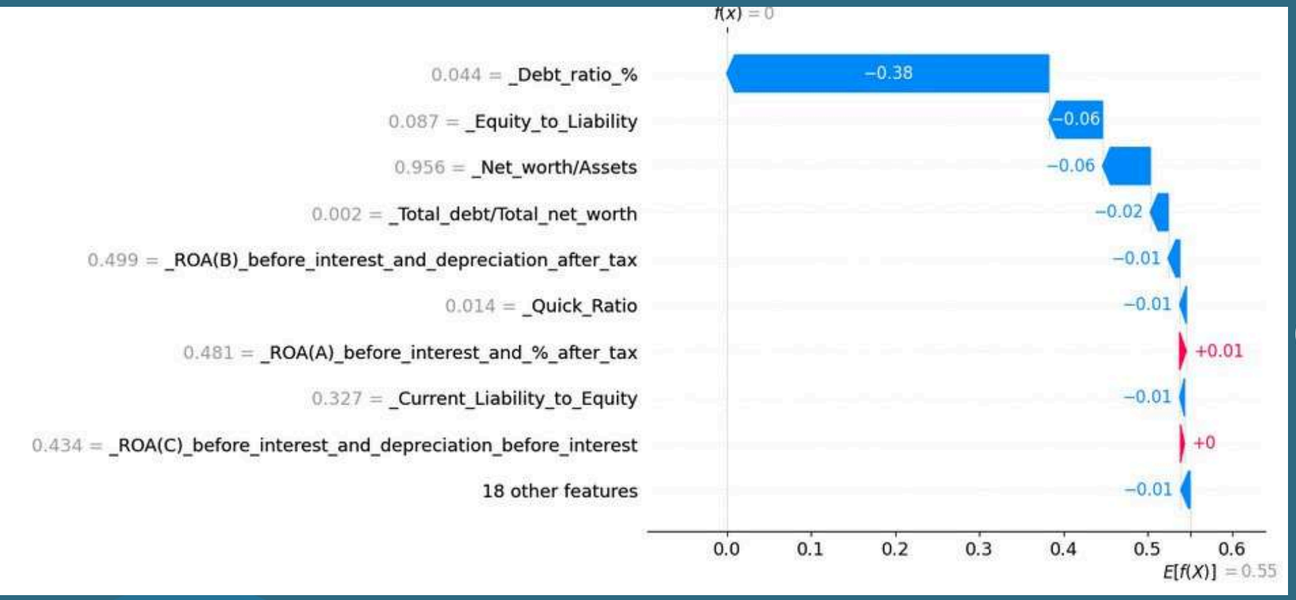
THE PARTIAL DEPENDENCE PLOT (PDP) ANALYSIS REVEALED THAT BOTH ROA BEFORE INTEREST AND DEPRECIATION, AND ROA AFTER INTEREST AND TAX, ARE STRONG, LINEAR PREDICTORS OF BANKRUPTCY RISK. AS EITHER ROA MEASURE INCREASES, THE PREDICTED PROBABILITY OF BANKRUPTCY STEADILY DECREASES IN A SMOOTH, MONOTONIC FASHION, WITH NO ABRUPT SHIFTS OR NON-LINEAR BEHAVIOR. THE MODEL IS MOST RELIABLE IN THE MID-RANGE (0.48–0.60), WHERE MOST DATA POINTS LIE, CONFIRMING THAT HIGHER PROFITABILITY CONSISTENTLY LOWERS BANKRUPTCY RISK IN A PREDICTABLE WAY.

GLOBAL SHAP



SHAP ANALYSIS REVEALED THAT IS THE MOST INFLUENTIAL FEATURE GLOBALLY, WITH A STRONG POSITIVE SHAP VALUE INDICATING ITS DOMINANT ROLE IN PREDICTING BANKRUPTCY. OTHER MAJOR CONTRIBUTORS INCLUDE MULTIPLE *ROA VARIANTS, REFLECTING A COMPANY’S PROFITABILITY, ALONG WITH FEATURES LIKE ALL HIGHLIGHTING THE MODEL'S RELIANCE ON FINANCIAL HEALTH INDICATORS.

Local SHAP

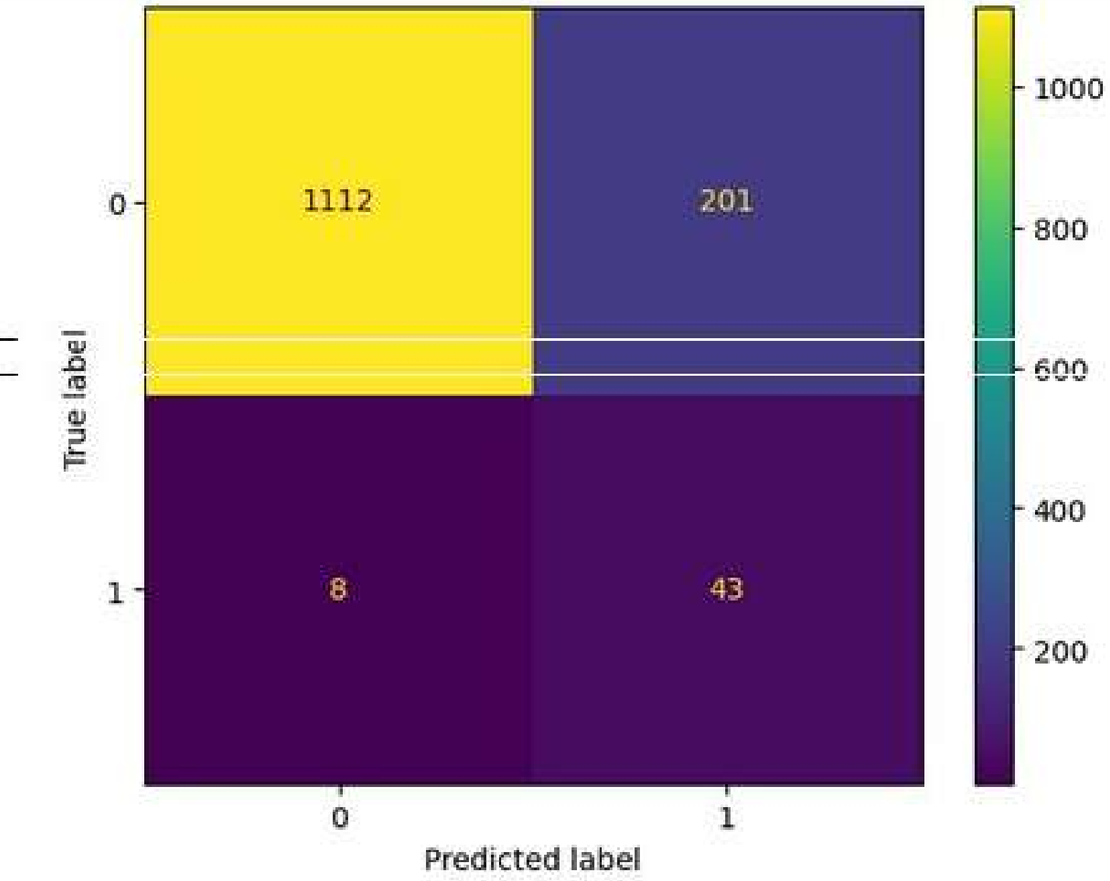


IN A LOCAL EXPLANATION FOR ONE COMPANY, DEBT RATIO% HAD THE LARGEST NEGATIVE SHAP IMPACT (.4), STRONGLY PUSHING THE PREDICTION TOWARD BANKRUPTCY. MINOR POSITIVE EFFECTS FROM FEATURES LIKE NET PROFIT BEFORE TAX/PAID-IN CAPITAL WEREN'T ENOUGH TO OFFSET THE RISK, INDICATING THAT THE PREDICTION WAS MAINLY DRIVEN BY HIGH DEBT AND WEAK FINANCIAL STABILITY.

MODEL 03

Classification Report:

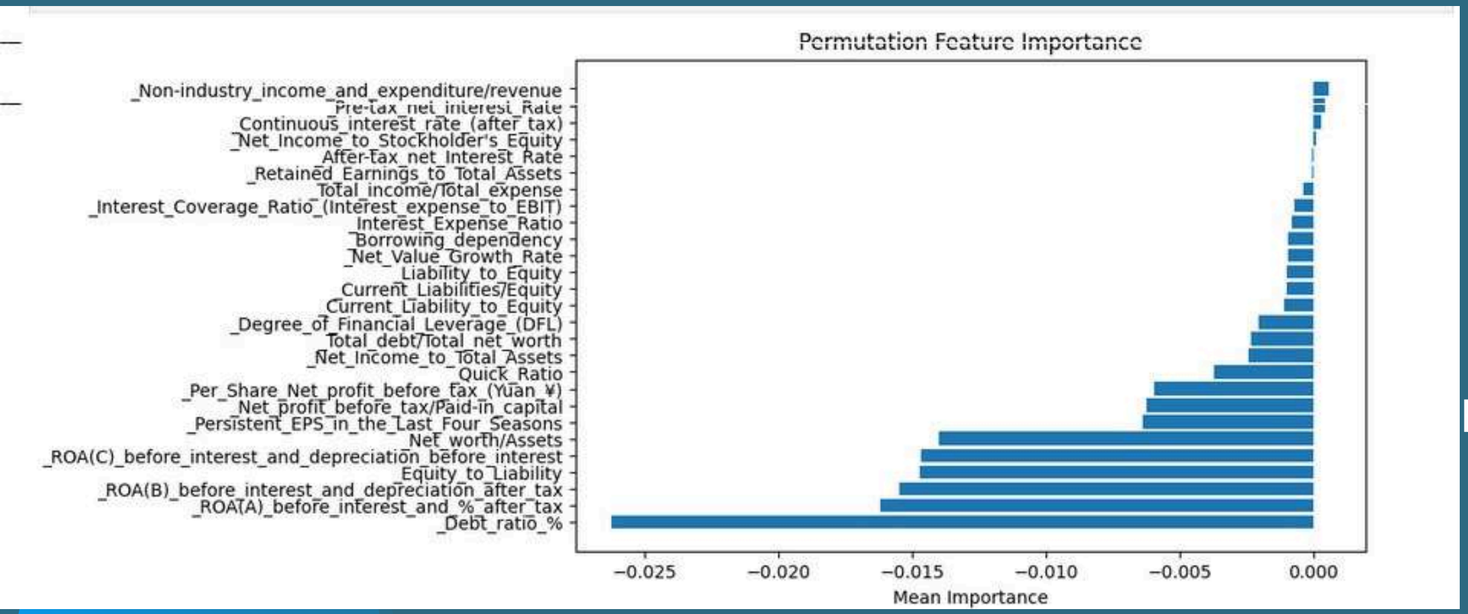
	precision	recall	f1-score	support
0	0.99	0.85	0.91	1313
1	0.18	0.84	0.29	51
accuracy			0.85	1364
macro avg	0.58	0.85	0.60	1364
weighted avg	0.96	0.85	0.89	1364



The neural network shows strong sensitivity to detecting bankrupt companies (high recall) but suffers from low precision, leading to many false alarms. It's effective in identifying risk but may benefit from further tuning or threshold adjustment to improve reliability in real-world scenarios.

EXPLAINABILITY TECHNIQUES LOGISTIC REGRESSION MODEL

Permutation Feature Importance



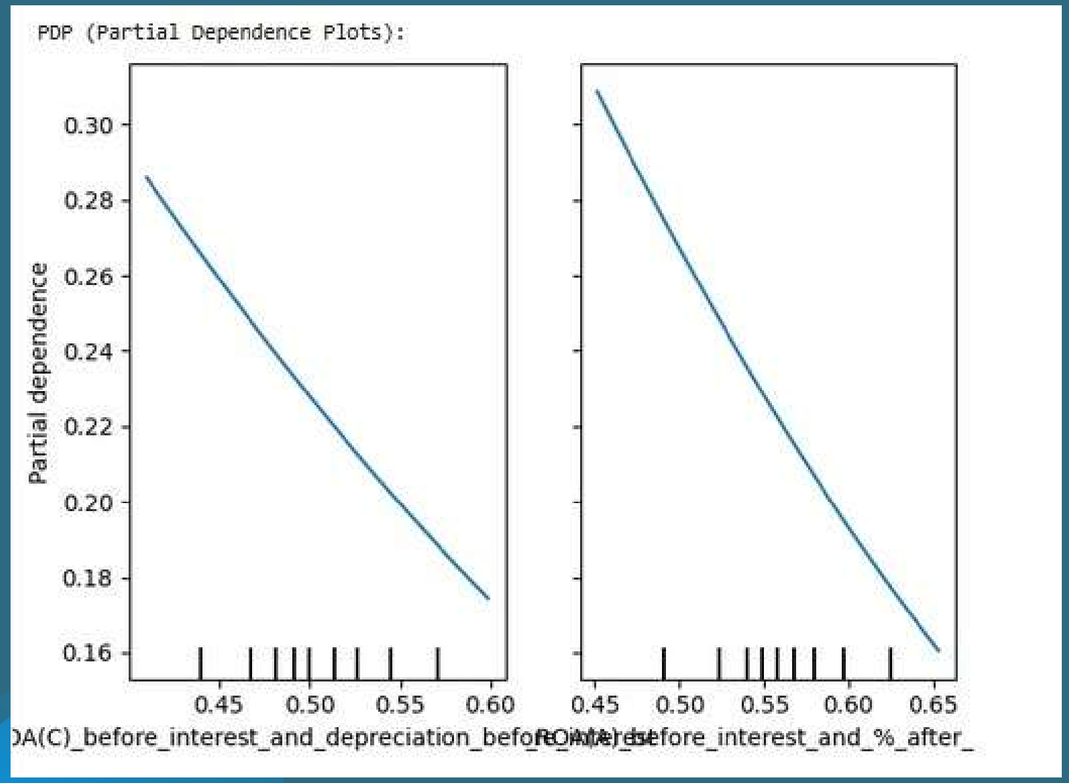
TOP PREDICTOR: DEBT RATIO HAD THE HIGHEST INFLUENCE, ALIGNING WITH SHAP RESULTS.

OTHER KEY FEATURES: ROA METRICS AND EQUITY TO LIABILITY ALSO SHOWED STRONG CONTRIBUTIONS.

LOW VALUE FEATURES: SOME VARIABLES HAD MINIMAL OR NEGATIVE IMPORTANCE, OFFERING LITTLE TO NO PREDICTIVE POWER.

CONCLUSION: PROFITABILITY AND LEVERAGE RELATED FEATURES ESPECIALLY DEBT RATIOARE CONSISTENTLY THE STRONGEST INDICATORS OF BANKRUPTCY RISK.N THE TOP 5 7 IMPACTFUL FEATURES.

PDP



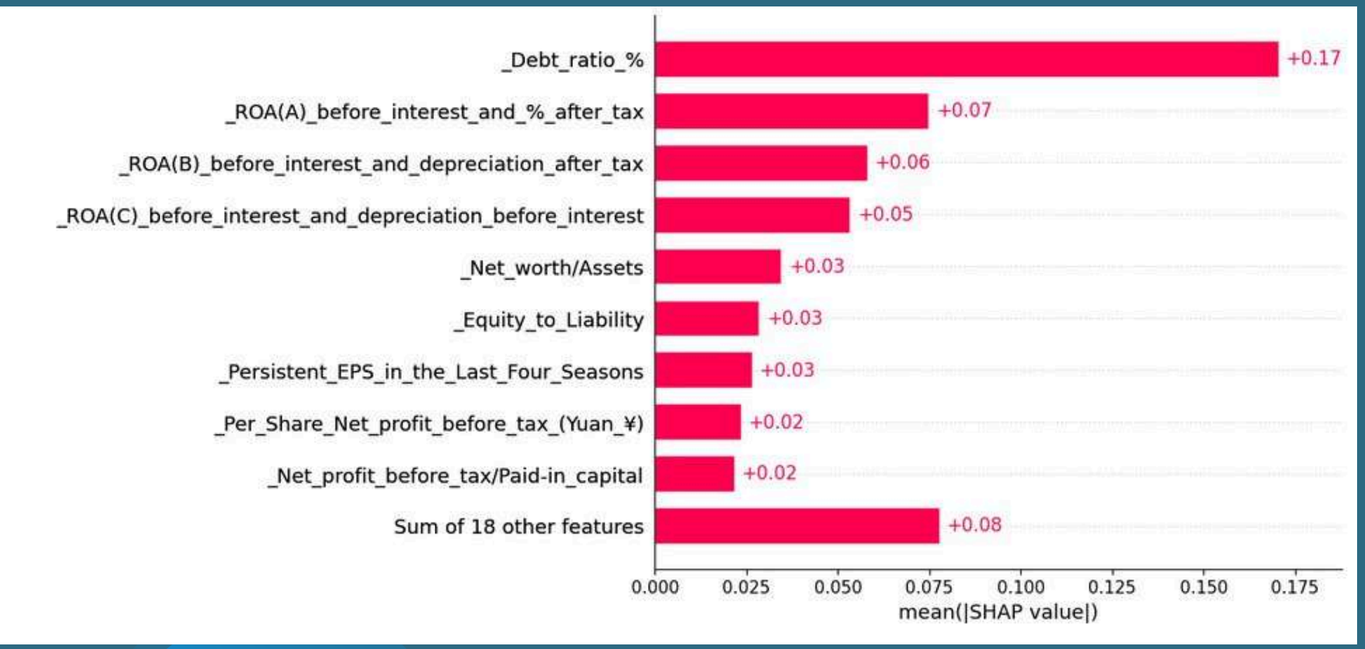
PDP RESULTS SHOW THAT ROA(C) AND ROA(A) HAVE A STRONG NEGATIVE RELATIONSHIP WITH BANKRUPTCY PROBABILITY MEANING AS THESE PROFITABILITY MEASURES INCREASE, THE LIKELIHOOD OF BANKRUPTCY SIGNIFICANTLY DECREASES. PROFITABILITY INDICATORS, PARTICULARLY ROA VARIANTS, PLAY A CRITICAL ROLE IN THE MODEL'S PREDICTIONS, REINFORCING THAT FINANCIALLY HEALTHY COMPANIES ARE LESS LIKELY TO GO BANKRUPT.

LIME

Feature	Value
_Quick_Ratio	0.01
_Net_Value_Growth_Rate	0.00
_Debt_ratio_%	0.04
_Equity_to_Liability	0.09
_Total_debt/Total_net_worth	0.00
_Net_worth/Assets	0.96
_Liability_to_Equity	0.28
_ROA(A)_before_interest_and_%_after_tax	0.48
_Interest_Expense_Ratio	0.63
_ROA(C)_before_interest_and_depreciation_before_interest	0.43

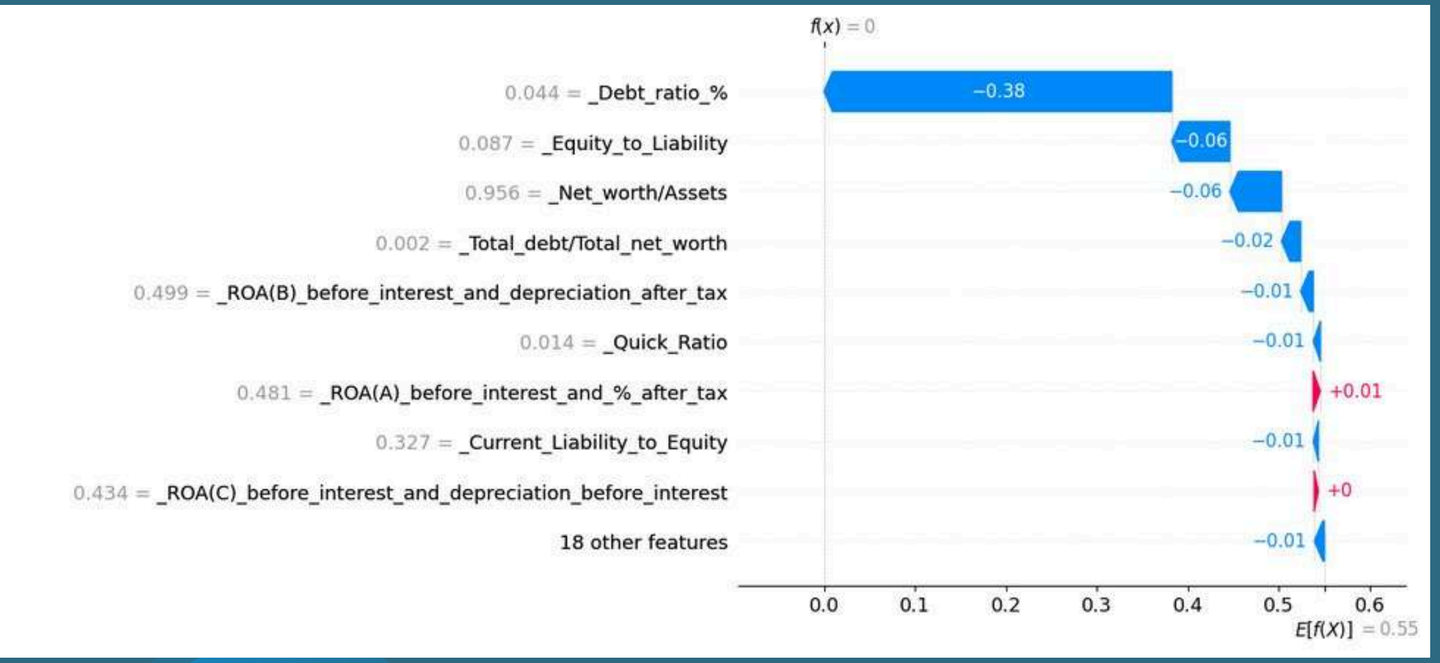
THE MODEL WAS 88% CONFIDENT THE COMPANY IS NOT BANKRUPT , MAINLY DUE TO A HIGH *ROA (0.50), LOW QUICK RATIO (0.01), AND AN EQUITY-TO-LIABILITY RATIO OF 0.09, ALL POINTING TO FINANCIAL STABILITY. LESS IMPACTFUL FEATURES LIKE NET VALUE GROWTH RATE AND NET PROFIT BEFORE TAX PLAYED ONLY A MINOR ROLE IN THE PREDICTION.

SHAP global



THE SHAP GLOBAL FEATURE IMPORTANCE ANALYSIS REVEALS THAT DEBT RATIO% IS THE MOST INFLUENTIAL PREDICTOR OF BANKRUPTCY, WITH A SHAP VALUE AROUND +0.17. OTHER KEY FEATURES CONTRIBUTING TO THE MODEL'S DECISIONS INCLUDE SEVERAL PROFITABILITY METRICS SUCH AS ROA(A), ROA(B), AND ROA(C), AS WELL AS FINANCIAL STABILITY INDICATORS LIKE NET WORTH/ASSETS, EQUITY TO LIABILITY, AND PERSISTENT EPS IN THE LAST FOUR SEASONS. OVERALL, THE MODEL PLACES SIGNIFICANT EMPHASIS ON BOTH DEBT LEVELS AND PROFITABILITY IN ASSESSING BANKRUPTCY RISK.

SHAP waterfall plot



THE SHAP WATERFALL PLOT, WHICH OFFERS A LOCAL EXPLANATION FOR A SINGLE COMPANY, SHOWS THAT DEBT_RATIO% HAD THE MOST SUBSTANTIAL NEGATIVE INFLUENCE (APPROXIMATELY –0.4) ON THE BANKRUPTCY PREDICTION. WHILE THERE WERE MINOR POSITIVE INFLUENCES, SUCH AS NET PROFIT BEFORE TAX/PAID-IN CAPITAL, THEY WERE NOT ENOUGH TO OFFSET THE IMPACT OF THE HIGH DEBT RATIO. THIS SPECIFIC PREDICTION HIGHLIGHTS HOW EXCESSIVE DEBT AND A WEAK FINANCIAL STRUCTURE CAN STRONGLY SWAY THE MODEL TOWARD FORECASTING BANKRUPTCY.

References

1. Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data | Computational Economics
2. (PDF) Review of bankruptcy prediction using machine learning and deep learning techniques
3. Bankruptcy Prediction Using Machine Learning Techniques
4. 2212.12051
5. 2401.12652
6. PERFORMANCE COMPARISON OF MULTIPLE DISCRIMINANT ANALYSIS AND LOGIT MODELS IN BANKRUPTCY PREDICTION – ProQuest
7. arxiv.org/pdf/2002.11705
8. An overview of bankruptcy prediction models for corporate firms: A Systematic literature review | Shi | Intangible Capital
9. Bankruptcy Prediction Using Machine Learning
10. Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession
11. Bankruptcy prediction using machine learning and Shapley additive explanations
12. Bankruptcy Prediction, Financial Distress and Corporate Life Cycle: Case Study of Central European Enterprises

Prepared by:
Nourhan Deif 202201959
Rahma Mourad 202201407
Radwa Badran 202200875
Rihana Nasr 202201092

THANK YOU