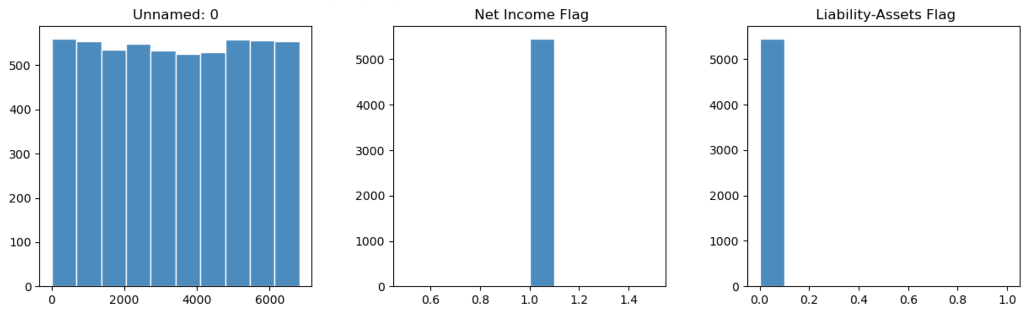
Company Bankruptcy Prediction Report – Zhiqi Duan

1. Data Preprocessing

Data Cleaning:

1. Irrelevant Columns Removal: Dropped “Unnamed:0” (company ID) and two binary flags (“Net Income Flag” and “Liability-Assets Flag”) due to low informational value.

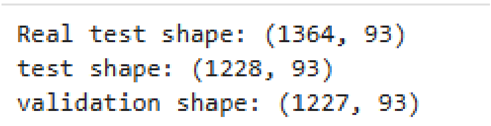


2. Missing Values: No missing values detected(traindata.isnull().sum() showed zero null entries)

Data Splitting Strategy:

1. To ensure the robustness and generalizability of the model, the dataset was divided into three parts: a training set, a validation set, and a test set. This structure allows for a comprehensive evaluation of the model's performance by assessing the consistency between the validation and test results. Consistent performance across these sets serves as an indicator of the model's robustness.

2. Additionally, the sizes of the validation and test sets were deliberately chosen to match the size of the actual test set for the final prediction task. This approach minimizes potential biases that could arise from variations in test set sizes, providing a more reliable assessment of the model's predictive capability.



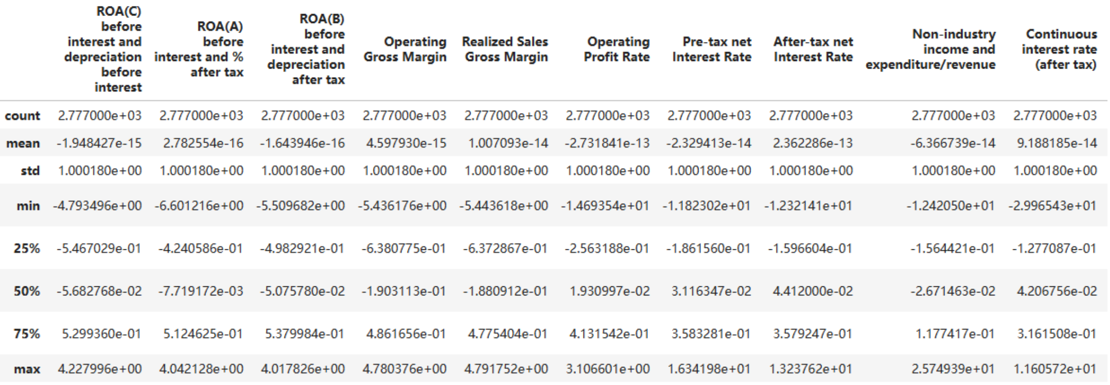
```
Real test shape: (1364, 93)
test shape: (1228, 93)
validation shape: (1227, 93)
```

Outlier Handling:

Winsorization: Since some models are particularly sensitive to the presence of outliers, a Winsorization technique was applied to mitigate their influence. Specifically, data points with values below the 0.1% percentile or above the 99.9% percentile were removed. This process eliminated 7.43% of extreme samples, reducing the training set from 3,000 to 2,777 observations. By reducing the impact of outliers, the model's stability and performance were further improved.

Data Standardization:

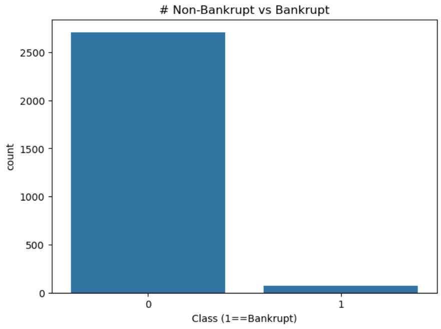
1. Applied StandardScaler to normalize numerical features, ensuring scale invariance for models. A comparison between RobustScaler and StandardScaler was conducted. Since outliers had already been effectively handled through Winsorization, StandardScaler was chosen as it demonstrated better overall performance.
2. To prevent data leakage and ensure a fair evaluation, the StandardScaler was fitted exclusively on the training set. The validation and test sets were then transformed using the same scaler without refitting. This approach maintained the integrity of the evaluation process and provided a more accurate assessment of the model's generalization ability.



	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	Continuous interest rate (after tax)
count	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03	2.777000e+03
mean	-1.948427e-15	2.782554e-16	-1.643946e-16	4.597930e-15	1.007093e-14	-2.731841e-13	-2.329413e-14	2.362286e-13	-6.366739e-14	9.188185e-14
std	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00	1.000180e+00
min	-4.793496e+00	-6.601216e+00	-5.509682e+00	-5.436176e+00	-5.443618e+00	-1.469354e+01	-1.182302e+01	-1.232141e+01	-1.242050e+01	-2.996543e+01
25%	-5.467029e-01	-4.240586e-01	-4.982921e-01	-6.380775e-01	-6.372867e-01	-2.563188e-01	-1.861560e-01	-1.596604e-01	-1.564421e-01	-1.277087e-01
50%	-5.682768e-02	-7.719172e-03	-5.075780e-02	-1.903113e-01	-1.880912e-01	1.930997e-02	3.116347e-02	4.412000e-02	-2.671463e-02	4.206756e-02
75%	5.299360e-01	5.124625e-01	5.379984e-01	4.861656e-01	4.775404e-01	4.131542e-01	3.583281e-01	3.579247e-01	1.177417e-01	3.161508e-01
max	4.227996e+00	4.042128e+00	4.017826e+00	4.780376e+00	4.791752e+00	3.106601e+00	1.634198e+01	1.323762e+01	2.574939e+01	1.160572e+01

Class Imbalance Mitigation:

After plotting the sample distribution, it was evident that the dataset was highly imbalanced, with a significantly smaller proportion of bankrupt companies compared to non-bankrupt ones. To address this, SMOTE-Tomek was applied, which combines Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links. This method oversamples the minority class (bankrupt) by generating synthetic examples and removes ambiguous samples using Tomek Links, effectively balancing the data.

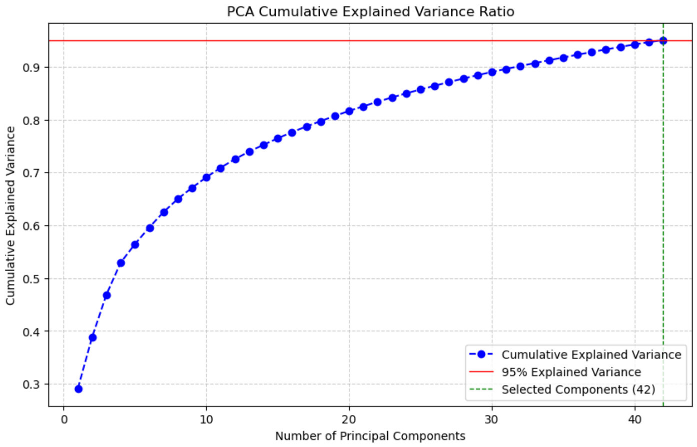


As a result, the training dataset expanded to 5,410 samples with an equal distribution of both classes. This balanced dataset helped mitigate the bias towards the majority class and improved the model's ability to accurately predict bankruptcy.

2. Feather Engineering

Dimensionality Reduction (PCA):

1. PCA with 95% Variance Retention: Reduced feature dimensions from 93 to 42.
2. Variance Explained: The first 42 principal components captured 95.06% of cumulative variance
3. However, based on the test results, models using PCA-transformed features consistently underperformed compared to those using the original feature set. As a result, PCA was ultimately not used for dimensionality reduction in the final model selection process.

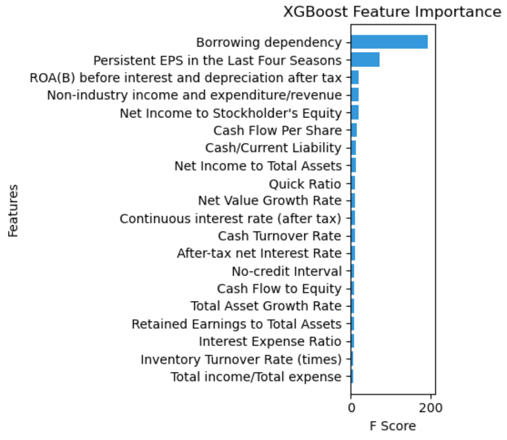


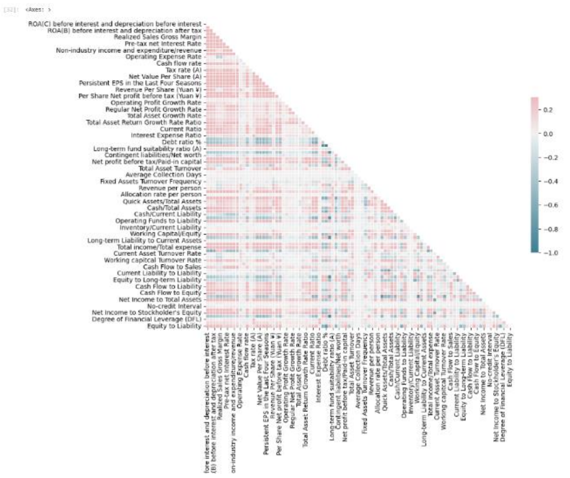
Feature Selection (XGBoost + correlation):

1. XGBoost Feature Importance: After training an XGBoost model on the processed data, features with importance scores greater than 0.001 were selected. Since there may be correlations among different columns, a further step was applied to mitigate potential overfitting.

Specifically, for pairs of features with a Pearson’s correlation coefficient exceeding 0.8, only the feature with the higher importance score was retained. This process reduced redundancy, ensured model simplicity, and improved generalization. Following this selection process, 53 features were retained.

2. Key Features: Borrowing dependency, Persistent EPS in the Last Four Seasons

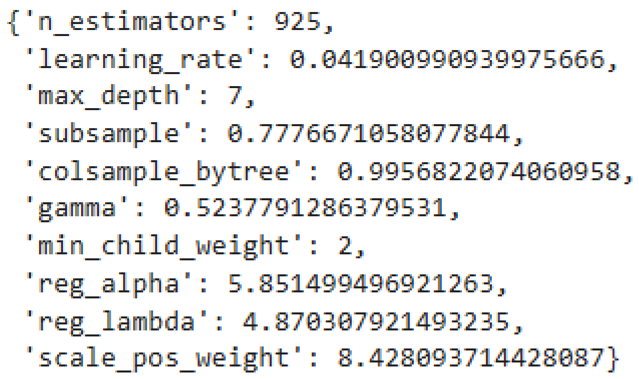




3. Model Building and Hyperparameter Tuning

XGBoost model (4 models):

1. Baseline Model: A baseline XGBoost model was initially trained using the default parameters (n_estimators=100, learning_rate=0.3) to establish a performance benchmark.
2. Feature-Based Model: Based on the results of the baseline model, the feature selection process described earlier was applied. Using the selected 53 features, a second XGBoost model was trained to evaluate the impact of feature reduction on model performance.
3. Hyperparameter Optimization: To further enhance performance, hyperparameter tuning was conducted using the Optuna framework. The optimization objective was to maximize the cross-validation average F1-score. A total of 30 Bayesian optimization trials were performed, searching within the following parameter space:



```
{'n_estimators': 925,
 'learning_rate': 0.041900990939975666,
 'max_depth': 7,
 'subsample': 0.7776671058077844,
 'colsample_bytree': 0.9956822074060958,
 'gamma': 0.5237791286379531,
 'min_child_weight': 2,
 'reg_alpha': 5.851499496921263,
 'reg_lambda': 4.870307921493235,
 'scale_pos_weight': 8.428093714428087}
```

The optimal hyperparameters identified through this process were used to train a third XGBoost model, leading to significant performance improvements.

4. Combined PCA and Optimized Model: A fourth model integrated PCA-based dimensionality reduction (42 components) with the optimized hyperparameters. Results showed no performance gain compared to the feature-selected and tuned model, likely due to information loss in PCA.

Support Vector Machines (2 models):

1. Linear SVM:
   Kernel: Linear kernel for efficiency and interpretability.
   Regularization: C=0.1 to balance margin width and misclassification penalties.
   Prioritized wider margins (C=0.1) to handle imbalanced data and reduce overfitting.
2. Polynomial SVM:
   Kernel: Degree-3 polynomial to capture nonlinear patterns.
   Pipeline: Combined Polynomial Features(degree=3) and StandardScaler for feature expansion and normalization.
   Tested nonlinear separability but faced convergence warnings, suggesting limited utility for this dataset.

Neural Network:

Layers:

Input: 128 neurons (ReLU) + Dropout (0.3).

Hidden: 64 neurons (ReLU) + Dropout (0.2).

Output: 1 neuron (Sigmoid).

Total Parameters: 7,361, balancing complexity and generalization.

Training Strategy:

Optimizer: Adam (learning_rate=0.001) with early stopping (patience=20) to prevent overfitting.

Loss Function: Binary cross-entropy for probabilistic predictions.

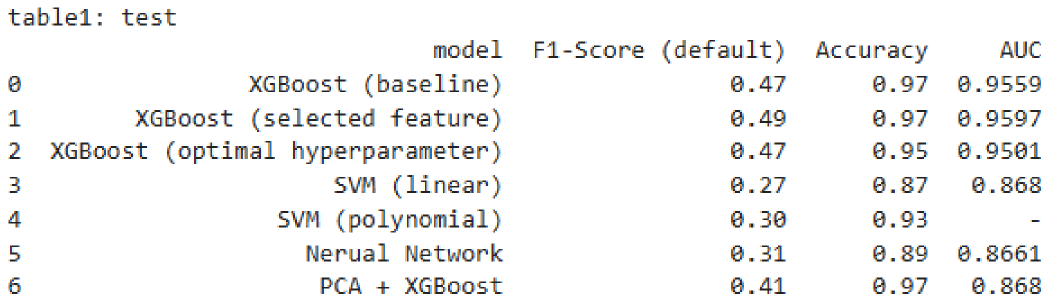
Class Imbalance: Relied on SMOTE-balanced data instead of explicit weighting.

Regularization

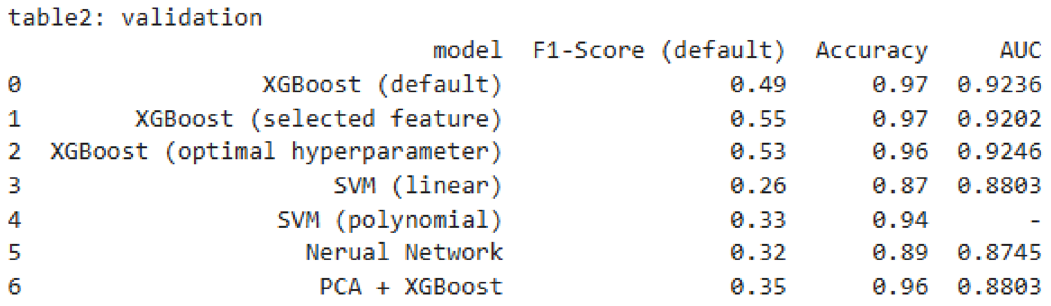
Dropout: Higher rate (0.3) in the input layer to handle noise, lower rate (0.2) in deeper layers to retain learned features.

4. Experimental Results and Comparison

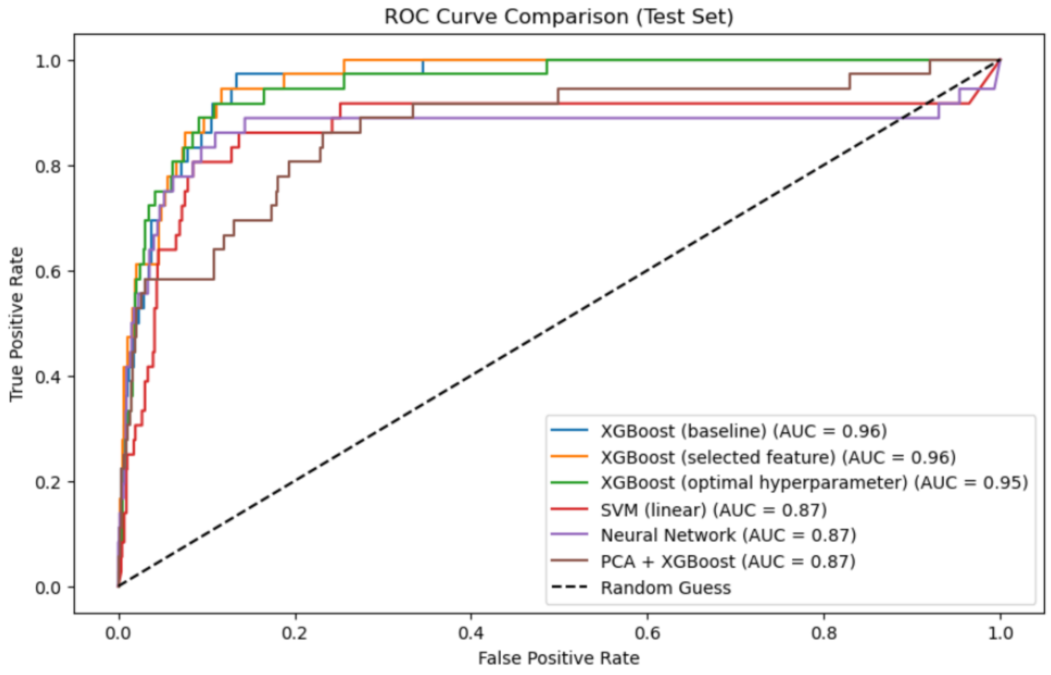
1. Compare the F1-Score, Accuracy, AUC, and ROC curve of my 7 models on both the test set and the validation set.



```
table1: test
                             model  F1-Score (default)  Accuracy     AUC
0                XGBoost (baseline)               0.47      0.97  0.9559
1          XGBoost (selected feature)             0.49      0.97  0.9597
2  XGBoost (optimal hyperparameter)               0.47      0.95  0.9501
3                      SVM (linear)               0.27      0.87   0.868
4                  SVM (polynomial)               0.30      0.93       -
5                   Nerual Network                0.31      0.89  0.8661
6                   PCA + XGBoost                 0.41      0.97   0.868
```

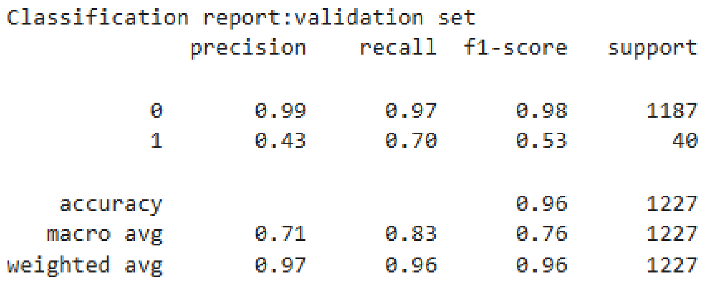


```
table2: validation
                             model  F1-Score (default)  Accuracy     AUC
0                 XGBoost (default)               0.49      0.97  0.9236
1          XGBoost (selected feature)             0.55      0.97  0.9202
2  XGBoost (optimal hyperparameter)               0.53      0.96  0.9246
3                      SVM (linear)               0.26      0.87  0.8803
4                  SVM (polynomial)               0.33      0.94       -
5                   Nerual Network                0.32      0.89  0.8745
6                   PCA + XGBoost                 0.35      0.96  0.8803
```



2. XGBoost with optimal hyperparameter and selected feature result (Best model) for the validation set:

Classification Report: F1-score=0.53

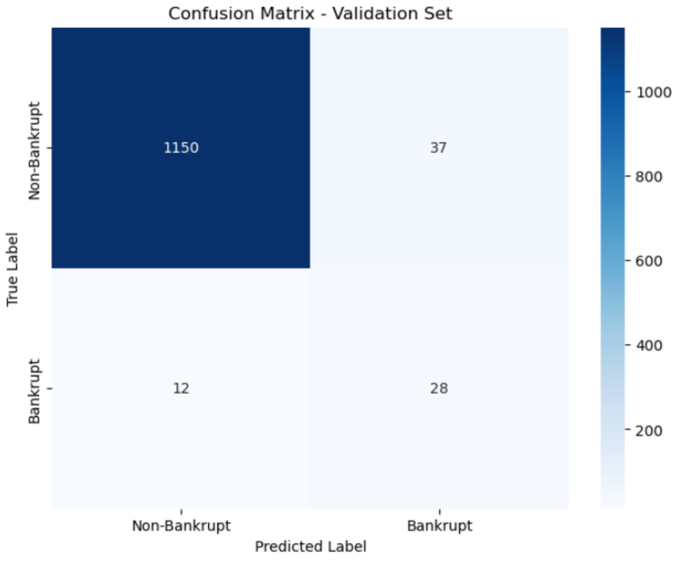


```
Classification report:validation set
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      1187
           1       0.43      0.70      0.53        40

    accuracy                           0.96      1227
   macro avg       0.71      0.83      0.76      1227
weighted avg       0.97      0.96      0.96      1227
```

Confusion Matrix:



5. Best Model Selection

The **optimized hyperparameter with selected feature XGBoost** was selected as the final model due to:

After performing five rounds of data splitting with different random seeds for the training, validation, and test sets, I aim to evaluate the performance of each model across these iterations. The objective is to assess their balance, consistency, robustness, and worst-case prediction performance. By synthesizing these insights, I will determine the most suitable model for selection.

Test Set F1 Score Times	XGBoost (baseline)	XGBoost (selected feature)	**XGBoost(optimal hyperparameter)**	PCA+XGBoost	SVM	Neural Network
1	0.39	0.38	**0.47**			
2	0.35	0.43	**0.38**	0.34	0.27	0.29
3	0.42	0.43	**0.40**	0.38		

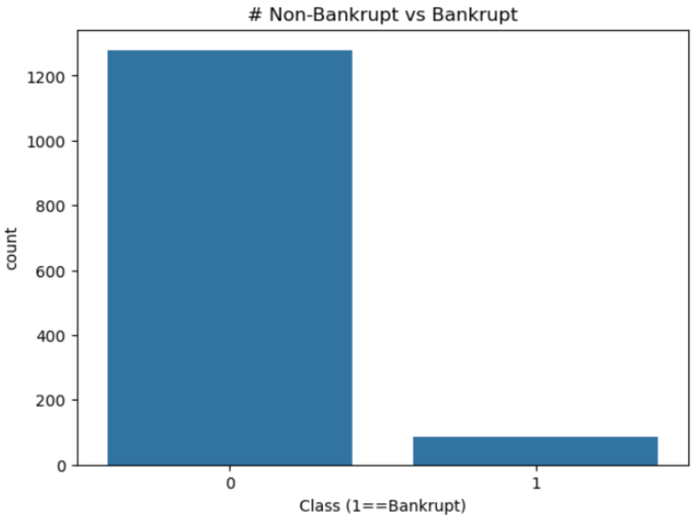
4	0.47	0.51	**0.46**	0.41		
5	0.25	0.29	**0.33**	0.29	0.22	0.28

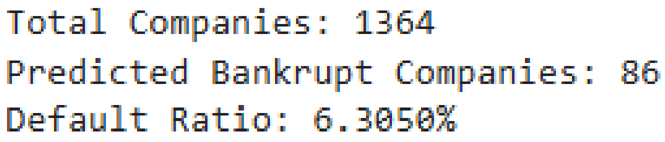
Validation Set F1 Score Times	XGBoost (baseline)	XGBoost (selected feature)	**XGBoost(optimal hyperparameter)**	PCA+XGBoost	SVM	Neural Network
1	0.25	0.31	**0.33**			
2	0.41	0.39	**0.36**	0.31	0.25	0.26
3	0.41	0.49	**0.43**	0.23		
4	0.49	0.47	**0.55**	0.35		
5	0.42	0.44	**0.47**	0.36	0.27	0.42

Comparing the results across the five iterations, the **XGBoost model with optimized hyperparameters and selected features** achieved the highest mean F1-score, the lowest variance, and the best worst-case performance. These results demonstrate the model's superior balance, consistency, robustness, and resilience in worst-case scenarios.

6. Test Set Prediction

Prediction of Bankrupt and Non-Bankrupt ratio with Best XGBoost model:




```
Total Companies: 1364
Predicted Bankrupt Companies: 86
Default Ratio: 6.3050%
```