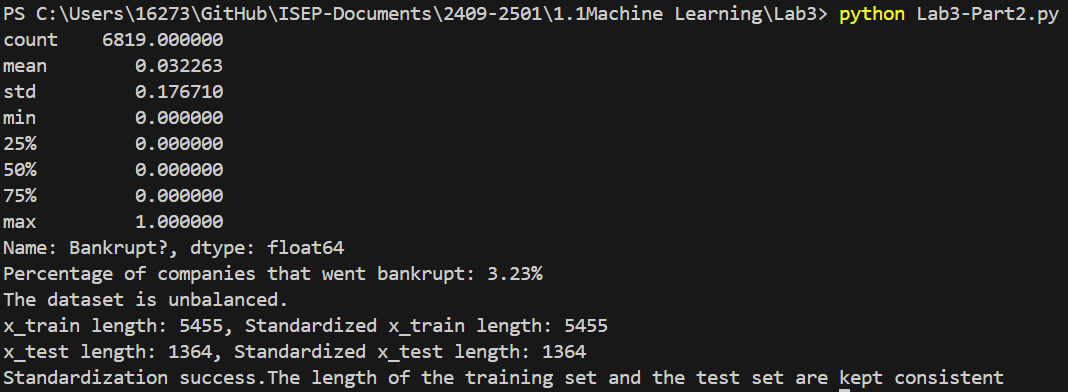
图片包含 文本

描述已自动生成

**GUO Xiaofan, FU Jintao**

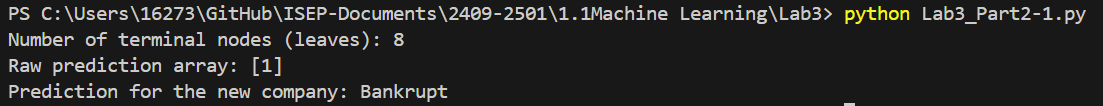
# **2. Part II: Practical applications**



1. Percentage of companies that went bankrupt: **3.23%**.
2. The dataset is **unbalanced**.
3. The length of the training set and the test set are **kept consistent**: x\_train length is 5455, x\_test length is 1364. Standardization success.

## **Classification Trees**

1.



图示

描述已自动生成

* The top-level nodes contain **segmentation features, Gini index, total number of samples, and category distribution information.**
* The Gini index is used to measure node purity, and the smaller the Gini index, the clearer the classification.
* The tree has **8** terminal nodes.
* The company **will go bankrupt**.

2.

图片包含 表格

描述已自动生成

**Training Set:**

* **Precision:** The prediction precision for bankrupt companies is 0.89, and for non-bankrupt companies is 0.97.
* **Recall:** For bankrupt companies, the recall is only 0.14 (i.e., the model only correctly identified 14% of bankrupt companies), while the recall for non-bankrupt companies is almost 1.00.
* **F1 score:** The F1 score for bankrupt companies is 0.24, which is low, indicating that the model does not have a good overall classification effect on bankrupt companies.
* **Accuracy:** The overall accuracy is 97%, indicating that the model performs well overall on the training set, but its ability to identify bankrupt companies is weak.

**Test Set:**

* **Precision:** The precision for bankrupt companies is low at 0.30, indicating that the model incorrectly predicts many non-bankrupt companies as bankrupt.
* **Recall:** The recall for bankrupt companies is 0.07 (only 7% of bankrupt companies are correctly predicted), indicating that the model has very poor recognition of bankrupt companies on the test set.
* **F1 score:** The F1 score for bankrupt companies is only 0.11, which is very low, indicating that the model is poor at classifying bankrupt companies.
* **Accuracy:** The overall accuracy is 96%, but the accuracy is affected by the high recall of non-bankrupt companies (which are the majority).

**Overfitting exists**: The model performs well on the training set, but its performance in classifying bankrupt companies on the test set drops sharply, which indicates that the model has overfit to the training data. It has learned too much detail or noise in the training data and cannot generalize to new data.

Solutions to Reduce Overfitting:

* **Reduce Model Complexity:**

Limit Tree Depth: Decrease the ‘max\_depth’ parameter of the decision tree to make the model less complex. You can try smaller values like 2 or adjust other parameters like ‘min\_samples\_split’ or ‘min\_samples\_leaf’ to control the growth of the tree.

* **Use Pruning Techniques:**

Cost Complexity Pruning: Use the ‘ccp\_alpha’ parameter to prune the decision tree after training. This will help remove parts of the tree that don't contribute much to the final prediction, reducing complexity and overfitting.

* **Cross-Validation:**

Apply cross-validation to find the best model parameters. This ensures the model is not overfitting the training data and can generalize better to new data.

* **Increase Training Data:**

Add more data, especially for bankrupt companies. More data can help the model learn better patterns and avoid overfitting to specific details in the training set.

c. 图表, 直方图

描述已自动生成

图形用户界面, 文本

描述已自动生成

* The best value of α: 0.00018331805682859762
* Number of leaves in the chosen tree: 105

d.

图片包含 文本

描述已自动生成

**Training Set:**

* **Confusion Matrix:** The model correctly predicted almost all companies, with only 2 misclassified as not bankrupt and 23 as bankrupt.
* Precision, Recall, F1-score:
  + For non-bankrupt companies (class 0), the scores are perfect (1.00).
  + For bankrupt companies (class 1), Precision is 0.99, Recall is 0.87, and F1-score is 0.92. The model performs well overall on the training set.

**Test Set:**

* **Confusion Matrix:** The model correctly predicted 1289 non-bankrupt companies but misclassified 29 as bankrupt. It correctly predicted 21 bankrupt companies but misclassified 25 as non-bankrupt.
* Precision, Recall, F1-score:
  + For non-bankrupt companies, Precision and Recall are both 0.98.
  + For bankrupt companies, Precision is 0.42, Recall is 0.46, and F1-score is 0.44. The model struggles more with predicting bankrupt companies on the test set.

The model performs well for non-bankrupt companies but has difficulty predicting bankrupt companies, especially on the test set.

## Ensemble methods: Bagging, Random Forest and Boosting

1.

图片包含 文本

描述已自动生成

**Training Set:**

* The model performs almost perfectly on the training set, indicating potential overfitting as it has learned the training data too well.

**Test Set:**

* The model performs well for non-bankrupt companies on the test set but struggles to correctly classify bankrupt companies, with a low recall of only 26%.

2.

文本

描述已自动生成

**Training Set:**

* The model performs perfectly on the training set with an accuracy of 1.00. Both precision and recall for bankrupt and non-bankrupt companies are very high, indicating overfitting to the training data.

**Test Set:**

* For non-bankrupt companies (class 0), precision and recall are still very high (0.97 and 0.99, respectively).
* For bankrupt companies (class 1), precision is 0.50, slightly better than Bagging, but recall is only 0.22, meaning the model misses most bankrupt companies.
* Overall accuracy is 0.97, similar to Bagging, but the model still struggles with identifying bankrupt companies.

**Comparison to Bagging:**

* Improvement: Random forest shows a small improvement in precision for bankrupt companies (0.50 compared to Bagging's 0.48).
* Limitations: The recall for bankrupt companies remains low (0.22), so the model still struggles to identify them.

3.

文本

描述已自动生成

**Training Set:**

* The model performs well for non-bankrupt companies (class 0), with high precision (0.98) and recall (0.99).
* For bankrupt companies (class 1), precision is low at 0.39, and recall is 0.28, meaning the model misses many bankrupt companies.
* Overall accuracy is 0.96.

**Test Set:**

* The model continues to perform well for non-bankrupt companies, with precision of 0.97 and recall of 0.98.
* For bankrupt companies, precision is 0.32 and recall is 0.22, meaning many bankrupt companies are misclassified.
* Overall accuracy is 0.96, but the model struggles with identifying bankrupt companies.

**Conclusion:**

* AdaBoost works well for non-bankrupt companies but struggles with bankrupt companies, especially in recall. Improving the model's ability to identify bankrupt companies is needed.

4.

图表, 条形图

描述已自动生成

图表, 条形图

描述已自动生成

图表, 条形图

描述已自动生成

文本

描述已自动生成

**Bagging:**

* The most important features are NITA, DR, and WKTA. NITA is the most significant (15.5%), followed by DR (13.3%). The importance is more balanced across the features.

**Random Forest:**

* Similar to Bagging, the most important features are DR and NITA, with ROAA also significant. It spreads the importance more evenly but focuses on these top features.

**AdaBoost:**

* Focuses almost entirely on NITA (50%), WKTA (25%), and DR (25%), ignoring all other features. It makes the model more interpretable but relies heavily on just a few key variables.

**Conclusion:**

* Across all classifiers, NITA, DR, and WKTA are consistently the most important predictors for bankruptcy.

## Support Vector Machines (SVM)

1.

文本

描述已自动生成

**Training Set:**

* The model performs very well with 99% accuracy.
* For non-bankrupt companies (class 0): Precision and recall are nearly perfect (99% and 100%).
* For bankrupt companies (class 1): Precision is high at 99%, but recall is only 65%, meaning some bankrupt companies are misclassified.

**Test Set:**

* The overall accuracy is still high at 96%.
* For non-bankrupt companies (class 0): Precision and recall are still excellent (98% and 99%).
* For bankrupt companies (class 1): Precision drops to 46%, and recall is only 28%, indicating many bankrupt companies are misclassified.

**Conclusion:**

* The model performs well for non-bankrupt companies but struggles to identify bankrupt companies, particularly in the test set. This suggests some overfitting, and the model could benefit from tuning to improve its recall for bankrupt companies.