VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FINAL REPORT

Topic:

PERFORMANCE PREDICTION OF ENTERPRISES IN VIETNAM BY RANDOM FOREST

Lecturers: **Assoc Prof PhD. Nguyen Anh Phong**
MBA. Phan Huy Tam

Summited by: **Nguyen Thi Hue Minh**
ID: **K194141733**

Ho Chi Minh City, June 23th, 2022

TABLE OF CONTENT

1. Introduction

1.1. Reason choose the topic

In today's competitive market, in order to survive and develop, enterprises need to find ways to increase profits in a reasonable way. To do so, enterprises need to improve their operational efficiency. The performance of the enterprises is also an investor's concern for the enterprises. The profitability of the business is a ratio to measure performance, this is a factor that helps managers develop businesses, and investors consider to decide to invest.

1.2. Literature review

1.2.1. Random forest

According to the definition of Breiman (2001): random forest is “a classifier consisting of a collection of tree-structured classifiers {h(x, Θk), k = 1, . . .} where the {k} are independently identically distributed random vectors and each tree casts a unit vote for the most popular class at input x”.

Random forest is an upgraded algorithm from decision tree classification. This algorithm allows to generate many random decision trees by using bootstrapping technique or randomly taking features. Then the trees will start to select and vote, the decision tree with the most votes will be the result of the model's prediction.

Random forest brings many advantages of decision tree and also overcomes some disadvantages of this algorithm. These advantages help to overcome the disadvantages of the data making the model better. Random forest robust with outliers, this algorithm can handle outliers contained in the data, like figure 3 and 4 mentioned below, the data used has a lot of outliers, and using random forest will deal with outliers this. Financial data is non-linear data, and random forests work well for this type of data. In addition, the algorithm also overcomes overfitting problem of decision tree and has higher accuraty than other classification algorithms.

1.2.2. Factors affecting enterprise’s performance

Size of the enterprise. According to Zeitun and Tian (2007) and Henrik Hansenand et al (2002), firm size has a positive effect on enterprise's performance.

Growth: high revenue growth of the business shows that the business is in the stage of increasing efficiency. Research by Zeitun and Tian (2007), enterprise's performance is positively related to revenue growth.

Fixed assets: these are the main assets of enterprises of great value, participating in many production and business cycles. Zeitun and Tian (2007), and Onaolapo and Kajola (2010)

show that the proportion of fixed assets has a negative impact on the performance of the enterprise.

Capital structure: The Pecking Order Theory by Meyers and Majluf (1984) shows a negative relationship between debt and profitability, i.e. Companies that use a lot of financial leverage will have lower profitability, which means low performance. The Trade-Off Theory of Capital Structure by Modigliani and Miller (1950), Zeitun and Tian (2007), and Onaolapo and Kajola (2010) show that the choice and use of capital affect the performance of the enterprise.

Liquidity: Businesses with the right level of liquidity can increase operational efficiency and reduce bankruptcy risk. However, when a business has too much liquidity, it may not be able to take advantage of all its resources and make assets redundant. This will increase the cost and reduce the performance of the enterprise.

2. Data

2.1. Data source

Data for the research are collected as panel data, include non-financial enterprise listed on the Ho Chi Minh City Stock Exchange and the Hanoi Stock Exchange. Data sources are supported by Thomson Reuters. Data is the figures from the annual financial statement of enterprises from 2009 to 2021.

2.2. Preprocessing data

First split the panel dataset by each data field. Next, convert the data such as: using the date row as a header, transposing data, format the data to float. To ensure that the data does not have too much missing value, delete companies with less than 5 years of establishment. Use interpolate() function to fill the missing value with the linear method. Because the column names of the data are in numeric format, change the column names to the names of the companies in the name variable.

2.3. Variables

Calculating the variables used for the model includes 11 features and 1 target. Features include, Size, PPE, Leverage, Liquidity, Growth, Days Sale Outstanding, Net profit margin, Total asset turnover, Cash holding, Debt to equity, Net working capital.

$$Size = \ln(Total\ asset)$$

$$PPE = \frac{Fixed\ asset}{Total\ asset}$$

$$Liquidity = \frac{Curent\ asset}{Current\ liabilities}$$

$$Growth = \frac{Revenue_t - Revenue_{t-1}}{Revenue_{t-1}}$$

$$Days\ sale\ outstanding = \frac{Receivable}{\frac{Sale}{365}}$$

$$Net\ profit\ margin = \frac{Net\ income}{Revenue}$$

$$Total\ asset\ turnover = \frac{Revenue}{Total\ asset}$$

$$Cash\ holding = \frac{Cash}{Total\ asset}$$

$$Debt\ to\ equity = \frac{Debt}{Equity}$$

$$Net\ working\ capital = \frac{Current\ asset - Current\ liabilities}{Total\ asset}$$

The Target variable is created from ROA and ROE with the condition:

- ROA >= 0.1 and ROE > 0.2: target = 1 (Effective business)

- ROA < 0.1 or ROE < 0.2: target = 0 (Ineffective business)

The target variable is created with the above condition because ROA and ROE are both ratios showing the profit-making performance of the business. These two metrics are considered good when ROA is greater than 7% and ROE is greater than 15%. However, because inflation in Vietnam is quite high, these two indexes are adjusted to ROA is greater than 10% and ROE is greater than 20% to match.

These variables are then put into a new dataframe with the index year and business name. The dataframe consists of 6144 rows and 12 columns. Check missing value of variables, no missing value.

3. Statistic descriptive and visualization

3.1. Statistic descriptive

	Size	PPE	Leverage	Liquidity	Growth	Days Sale Outstanding	Net profit margin	Total asset turnover	Cash holding	Debt to equity	Net working capital
count	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000	6144.000000
mean	27.193043	0.247771	0.217523	2.492739	6.714729	350.526422	0.052748	1.192837	0.063310	0.762913	0.224419
std	1.579222	0.219419	0.186861	5.744780	475.217809	10619.525264	2.090806	1.284589	0.092147	1.267551	0.224670
min	20.974620	0.000000	0.000000	0.000000	-4.515765	-40610.161558	-123.024748	-0.047302	-0.003437	-1.105085	-0.544713
25%	26.134743	0.074115	0.041160	1.133920	-0.178652	34.938549	0.020292	0.437389	0.013698	0.065317	0.066223
50%	27.125122	0.181403	0.187938	1.516974	-0.041412	72.553760	0.053426	0.889251	0.034207	0.395935	0.199904
75%	28.142255	0.357496	0.351932	2.404122	0.155876	143.705201	0.123461	1.504465	0.074290	1.041591	0.369649
max	33.691042	1.122133	0.954313	242.577244	37244.404527	818071.607608	28.473115	19.836378	1.741029	40.341193	0.986337

Figure 1. Statistic descriptive of features

As shown above, it can be seen that the variables PPE, Leverage and Liquidity have values > 0. The variables Leverage and Net working capital are all less than 1. The variables Leverage, Liquidity, Growth and Days sale outstanding have an unusually large max value. , and much larger than mean value, it is possible that these variables appear outlier. Standard deviation of Growth and Days sale outstanding is very high, showing that these two variables have strong volatility.

3.2. Visualization

Figure 2. Correlation

In general, most variables have low correlations with each other. Net working capital is highly correlated with PPE, Leverage, Debt to equity, and Liquidity. Leverage is highly correlated with Debt to equity and Size.

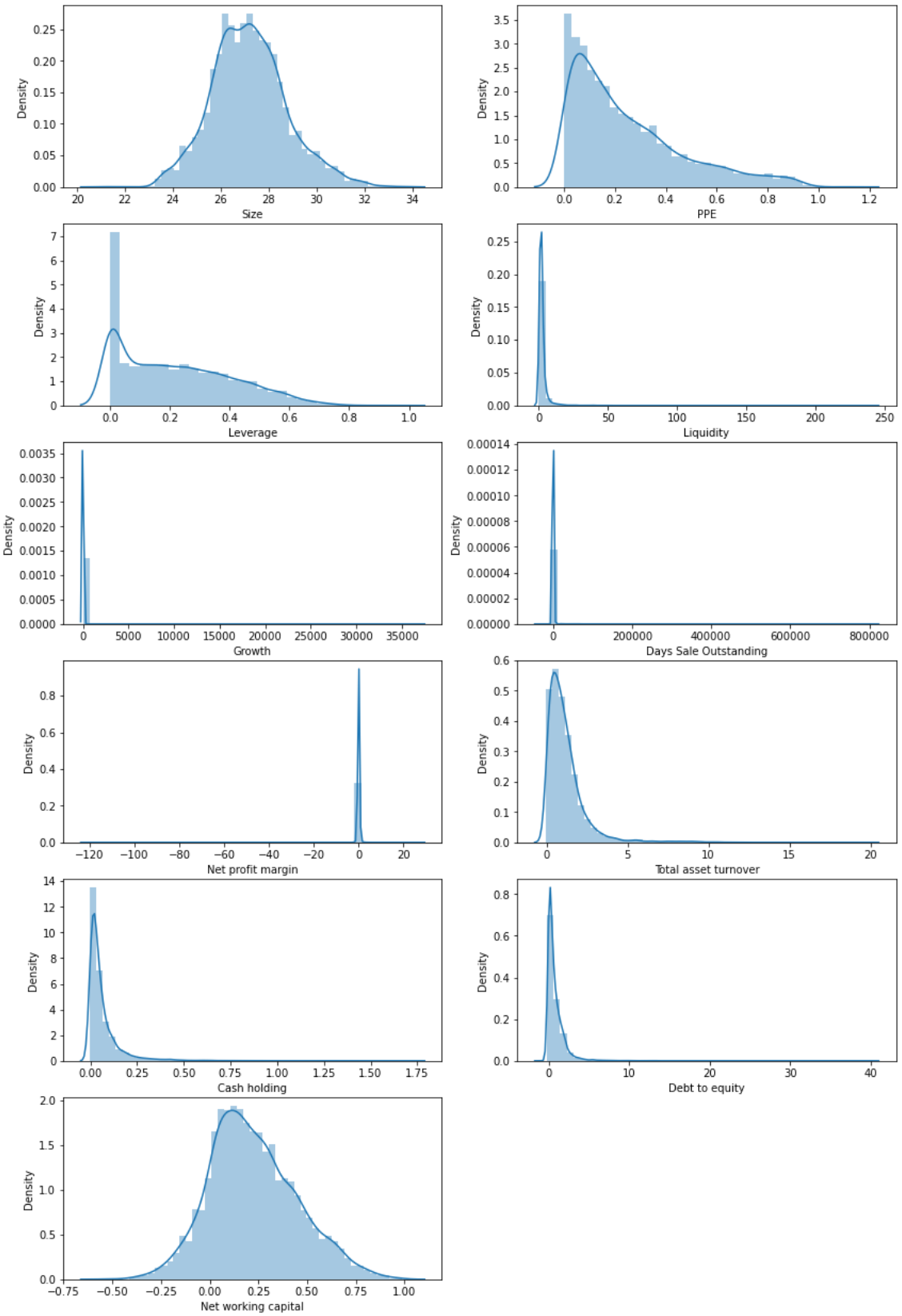

Figure 3. Distribution plot

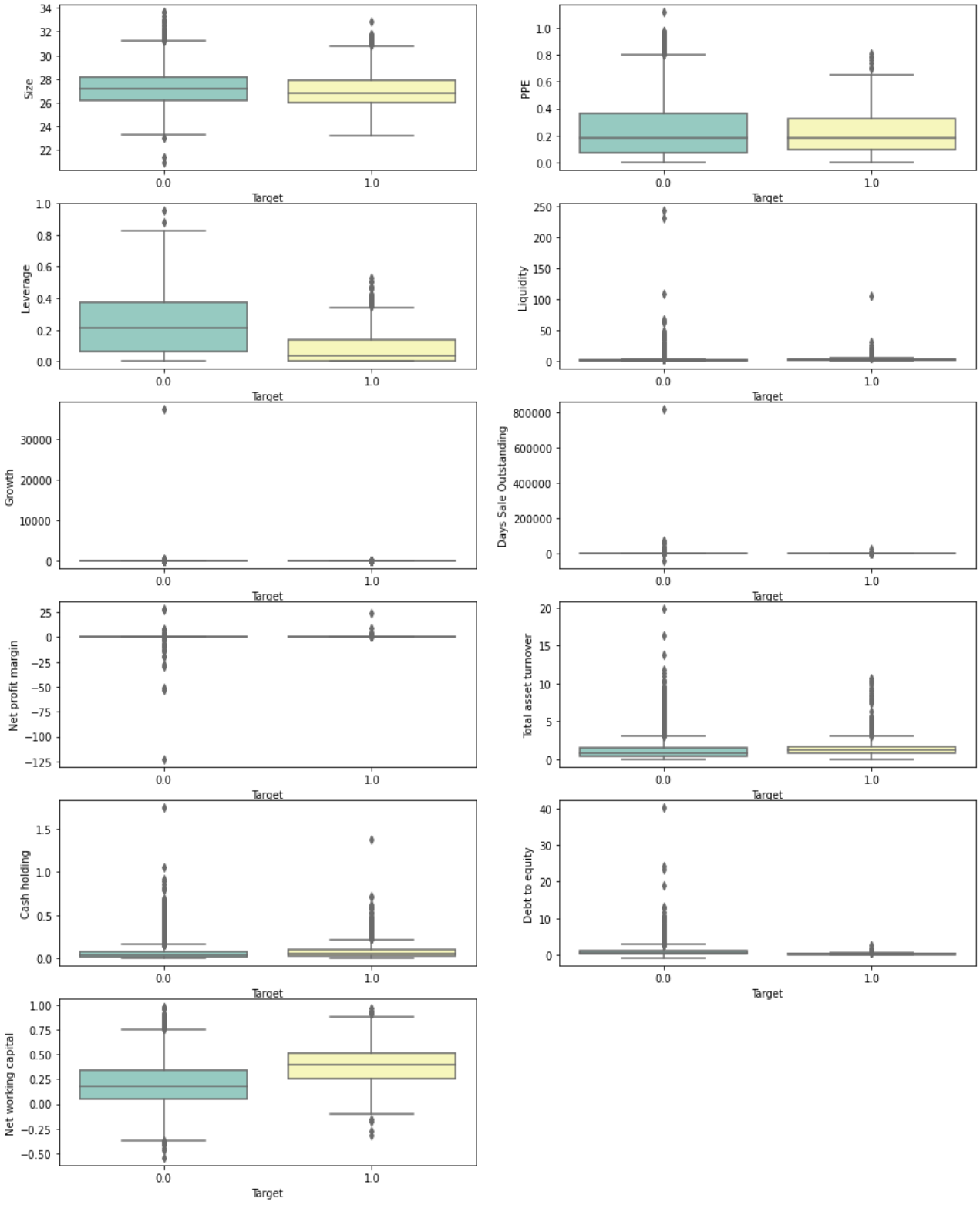

Figure 4. Box plots

The figures above show, 11 feature variables of the data all have outliers. The variables Liquidity, Growth, Days sale outstanding, Net profit margin, Total asset turnover, Cash holding, and Debt to equity have very large outliers, which causes the distribution plot and box plot of these variables to shrink. Comparing between classes 0 and 1, it can be seen that businesses with effective operations use less financial leverage because the min, mean, and max values of class 1 are smaller than those of 0. In contrast, these enterprises have higher net working capital than inefficient enterprises.

4. Model

4.1. Build model with training data

To train the predictive model with good results, I find for the parameter n_estimator - the number of decision trees most suitable for the random forest. When the number of decision trees is 91, the model can get an accuracy score of 96.61% as shown below.

```
optimal n_estimators value is 91 that accuracy is 0.966145114139034
```

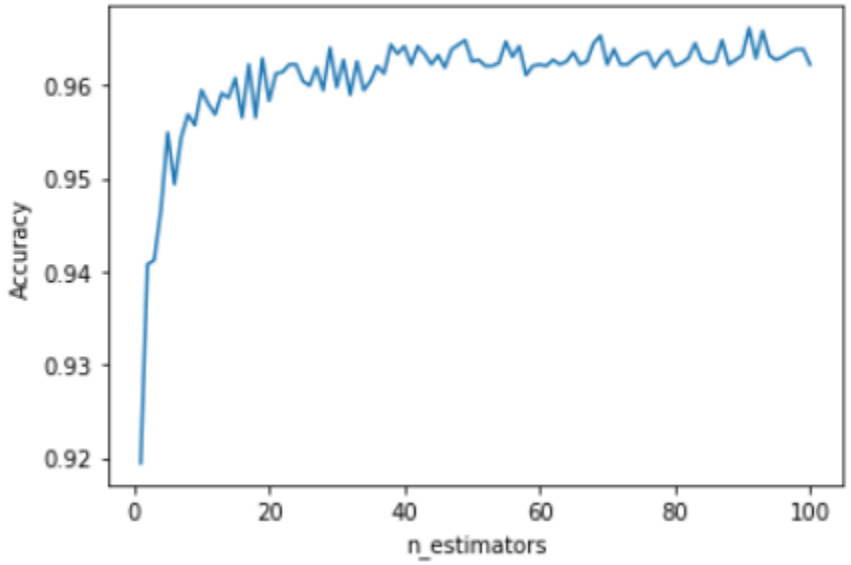

Figure 5. Accuracy line with the number of n_estimators

4.2. Oversampling data

Since the data is imbalanced, there are 5500 observations in class 0 and 644 observations in class 1, the training model may predict class 1 incorrectly. However, because the prediction results of class 1 are more important than the results of class 0, this is the class that is more interested. Therefore, it is necessary to overcome data bias to increase the accuracy of class 1 prediction. Due to oversampling data, increase the size of class 1 – less class. After upsampling class 1, the data is balanced: there are 5500 observations of class 0 and 5500 observations of class 1.

```
Target
0.0    5500
1.0    5500
dtype: int64
```

5. Result and conclusion

5.1. Result

5.1.1. Confusion matrix

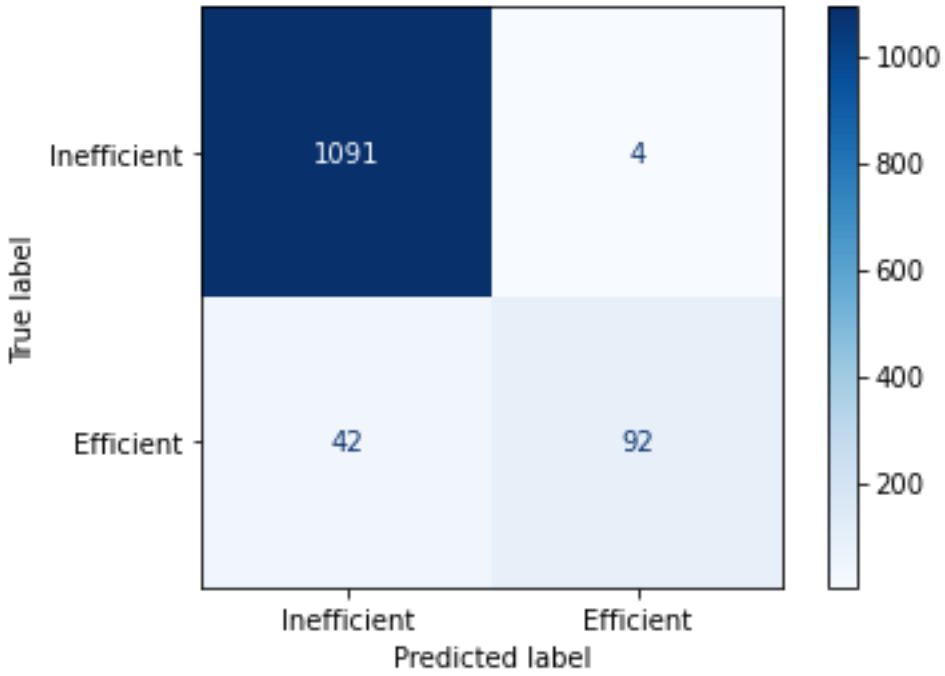

Figure 6. Confusion matrix

- True Positive = 1091: The model correctly predicted 1091 cases where businesses did not operate efficiently.
- False Positive = 42: There are 42 cases where businesses operate effectively but are wrongly classified as ineffective by the model.
- False negative = 4: There are 4 cases where the business is inefficient but is wrongly classified as effective by the model.
- True negative = 92: There are 92 cases of effective businesses that are correctly classified by the model

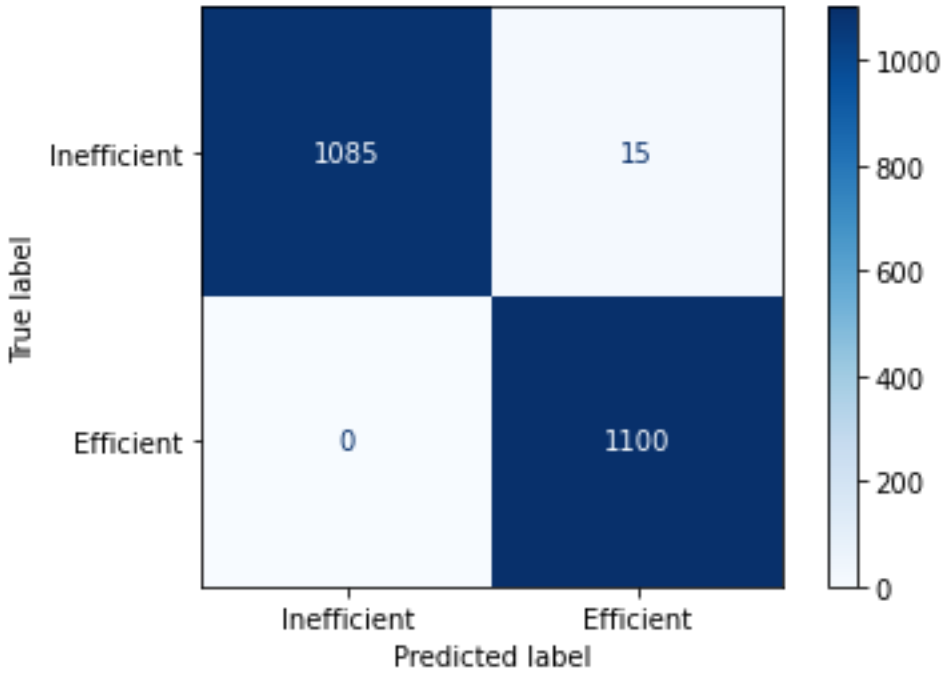

Figure 7. Confusion matrix with balanced data

- True Positive = 1085: The model correctly predicted 1085 cases the business did not operate efficiently.
- False Positive = 0 : There are no cases where effective businesses are misclassified.
- False negative = 15: There are 15 cases where the business is inefficient but is wrongly classified as effective by the model.
- True negative = 1100: All effective business cases have been correctly classified by the model.

5.1.2. Classification report

	precision	recall	f1-score	support
0.0	0.96	1.00	0.98	1095
1.0	0.96	0.69	0.80	134
accuracy			0.96	1229
macro avg	0.96	0.84	0.89	1229
weighted avg	0.96	0.96	0.96	1229

Random forest accuracy: 0.9625711960943857

Figure 8. Classification report

- Support0= 1095, support1= 134 -> Imbalanced dataset. This makes the metrics in class 1 are lower than class 0
- Recall of class 1 is much lower than that of class 0, this is because the data is imbalanced, and the number of class 1 is much less than that of class 0.
- The imbalanced dataset also makes the recall, f1-score calculated by the macro average are low, 84% and 89%. When weighted averages these metrics are pretty high 96%.
- Accuracy of the model is 96.26%, but due to the imbalanced dataset, it is impossible to trust the accuracy too much. F1-score of the macro average of 89% will be considered as the accuracy of the model instead of accuracy.

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	1100
1.0	0.99	1.00	0.99	1100
accuracy			0.99	2200
macro avg	0.99	0.99	0.99	2200
weighted avg	0.99	0.99	0.99	2200

Random forest accuracy: 0.9931818181818182

Figure 9. Classification report with balanced data

Sau khi upsample data: all metrics are increase (except for recall of class 0), all metrics are extremely good.

- Support 0 = support 1: data is balanced
- Accuracy increased from 96.26% to 99.32%
- Recall and f1-score after data balance also increase.
- Macro average and weighted average are equal.

5.1.3. Feature important

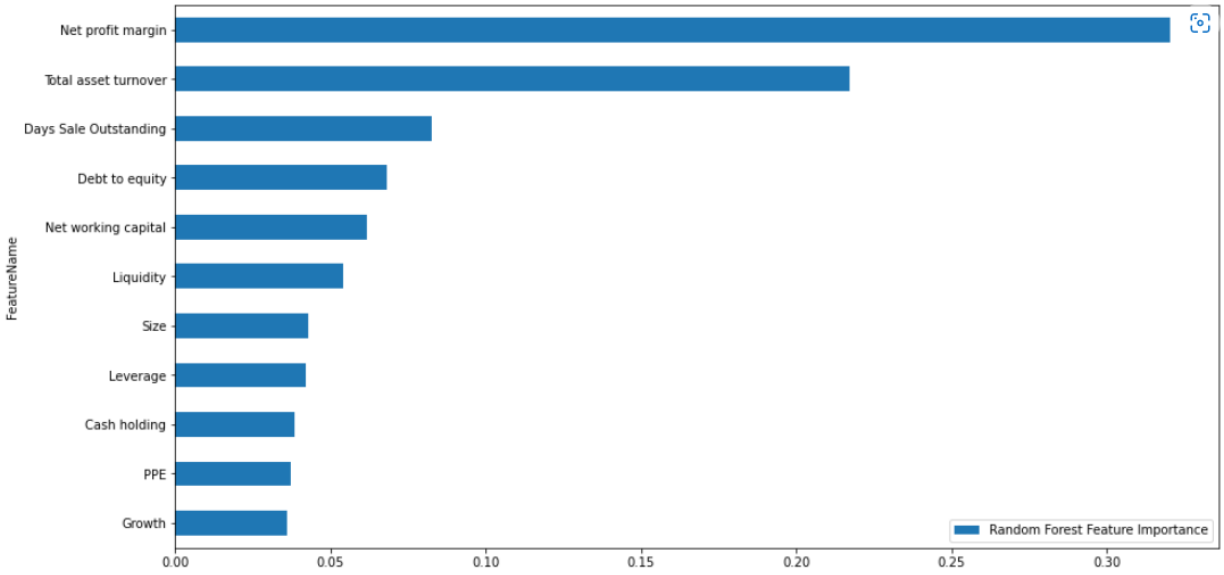

Figure 10. Feature importance

Based on the image above, it can be seen that the most important feature affecting the performance of the enterprise is net profit margin. The next 2 important features are total asset turnover, and days sale outstanding. Growth is the least important feature.

5.2. Conclusion

The study uses two profitability ratios, ROA and ROE, to classify the performance of a business. The random forest algorithm is used to produce a model with high accuracy and good prediction of class 1 (efficient). The algorithm handles outliers, making the model unaffected by them. The precision of class 1 in the model is the metric of most interest, which scores very well in both models with imbalanced and balanced data of 96% and 99%. The model also shows that the variables that have the most influence on business performance are Net profit margin, Total asset turnover, and Days sale outstanding. Businesses need to pay attention and adjust these ratios accordingly to improve performance. Although the classification model has good accuracy, it does not mean that the model will have good accuracy with other data sets. Because the variables used for the model are incomplete, macro variables such as inflation and interest rates or other financial indicators are missing. For a better predictive model, it is possible to increase the amount of data and other variables in the future.

REFERENCES

Breiman, L. 2001. *Machine Learning*, 45(1), 5-32.

Henrik Hansen, John Rand và Finn Tar (2002), SME Growth and Survival in Vietnam: Did Direct Government Support Matter?

Onaolapo, A and Kajola, S. (2010). "Capital Structure and Firm Performance: Evidence from Nigeria" European Journal of Economics, Finance and Administrative Sciences.

Zeitun, R. and Tian, G. G., Capital structure and corporate performance: evidence from Jordan, Australasian Accounting Business and Finance Journal, 1(4), 2007.