CLASSIFY FIRMS. FAST GROWTH OR NOT

DATA

The firms are classified based on their return on change from 2014 to 2015. Therefore, the data is filtered for the years 2014 and 2015. It consists of 39,770 firms and 49 variables. Only active firms are included in the analysis.

The firms that are above the 75th percentile of income to sales ratio are regarded as fast-growing firms.

FEATURE ENGINEERING

The metrics derived for the classification are

- current ratio
- debt to equity
- quick ratio
- return on sales (the change from 2014 4to 2015 determines the fast growth)
- fixed asset turnover

MODELS

Model Variables

Model variables consist of changes in certain variables from 2014 to 2015. They are 'inventory_c', 'tang_assets_c', 'intang_assets_c', 'current_ratio_c', 'debt_to_equity_c', 'quick_ratio_c', 'return_on_sales_c', 'fixed_asset_turnover_c'.

All models are built using the same variables.

Cost Function

False positives are considered as 2 times more costly than false negatives Because, from an investing point of view, investing in a firm that is not growing can be more costly.

*cost = fp*2 + fn*

RANDOM FOREST

Random forest is 5-fold cross-validated and built using random grid search. The best parameter are

- number of estimators: 1000
- maximum features: auto
- maximum depths: 140

The random model performs with 100% accuracy. Therefore, it does not need any adjustments for threshold.

LOGIT

The best logit model resulted in cost 21 with a 0 false positive rate. The best threshold for classification is 0.87.

Best Thresholds and FP, FN Cost Training Set

precision	recall	threshold	cost	tn	fp	fn	tp

1	0.99	0.877	21	11941	0	21	3935

LOGIT-LASSO

The best logit model with 5-fold cross-validation resulted in a cost 108 with a 1 false positive. The best threshold for classification is 0.508

Best Thresholds and FP, FN Cost Training Set

precision	recall	threshold	cost	tn	fp	fn	tp
0.99	0.97	0.508	108	11940	1	106	3850

CONCLUSION

We see that the random forest model performs the best with the highest accuracy and lowest cost.

Model Comparision Graph

	Random Forest	**Logit**	**Logit-Lasso**
FP	0	0	1
FN	0	21	106
Cost	0	21	108
Threshold	-	0.877	0.508

Therefore, the selection is the random forest model.